

**A COMPARATIVE EVALUATION OF THREE SITUATIONAL JUDGMENT
TEST RESPONSE FORMATS: ADDITIONAL ASSESSMENT OF
CONSTRUCT-RELATED VALIDITY, SUBGROUP DIFFERENCES,
SUSCEPTIBILITY TO RESPONSE DISTORTION, TEST-TAKER REACTIONS,
AND INTERNAL PSYCHOMETRIC PROPERTIES**

A Thesis

by

CRAIG DOUGLAS WHITE

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Chair of Committee,
Committee Members,

Head of Department,

Winfred Arthur, Jr.
Stephanie Payne
David Martin
Douglas Woods

May 2014

Major Subject: Psychology

Copyright 2014 Craig Douglas White

ABSTRACT

The primary objective of the present study was to investigate the construct-related validity of three situational judgment test (SJT) response formats. The present study addressed a potential common method bias threat arising from a shared-common-response-method effect that serves as a plausible alternative explanation for the posited differences-in-g-loading (i.e., more demanding in terms of information processing) explanation, which suggests that the variance in scores is due to the differences in the cognitive and information-processing demands of each response format. Thus, the present study used a design in which the three integrity-based SJT response formats were crossed with a GMA test and a personality measure using the same response formats, and 492 undergraduate students were randomly assigned to one of nine study conditions associated with the possible combinations of response formats. White–Hispanic and sex-based subgroup differences, susceptibility to response distortion, and test-taker reactions concerning the three response formats were also assessed, along with a comparative assessment of the internal consistency, test-retest, and alternate-form reliabilities of the three response formats. The results of this study generally supported the differences-in-g-loading explanation for the observed effects. In addition, mixed results were obtained for the differences in SJT scores between Whites and Hispanics, although the only significant difference was found for the rate response format, which favored White respondents. Consonant with the construct assessed by the SJT, women outperformed men on all three response formats, particularly the rank-SJT. The results

indicated that the relationship between response format and response distortion was strongest for the rate-SJT, followed by the rank- and then the most/least-SJTs.

Participants displayed the most favorable reactions to the most/least-SJT. Lastly, the internal consistency and test-retest reliability estimates were highest for the rate-SJT, while the alternate-form reliability estimates were highest for the rank- and most/least-SJTs. In summation, in the context of noncognitive constructs (e.g., integrity), the rate-SJT appears to be the superior, preferred response format, with its main drawback being its susceptibility to response distortion.

DEDICATION

To my parents, Doug and Barbara White, for their unconditional love and endless support in helping me reach my goals.

ACKNOWLEDGEMENTS

I would like to thank my committee chair, Dr. Winfred Arthur, Jr., and my committee members, Dr. Stephanie Payne and Dr. David Martin, for their guidance through the development and execution of this thesis project.

Thanks also go to my friends, colleagues, and the department faculty and staff for providing me with an exceptional environment in which to learn and grow as an industrial/organizational psychologist. I extend my gratitude to the Texas A&M University Office of Graduate Studies for awarding me with a Merit Fellowship, which has funded my research and coursework, as well as the undergraduate students who participated in this project.

Lastly, great thanks go to my parents, wife, and children for their patience and support.

TABLE OF CONTENTS

	Page
ABSTRACT.....	ii
DEDICATION.....	iv
ACKNOWLEDGEMENTS.....	v
LIST OF TABLES.....	viii
LIST OF FIGURES.....	x
INTRODUCTION AND LITERATURE REVIEW.....	1
Strengths of Situational Judgment Tests.....	1
Situational Judgment Tests in Practice.....	2
Methods vs. Constructs.....	3
Initial Research on Situational Judgment Test Response Formats (Glaze, Jarrett, Arthur, Schurig, & Taylor, 2011).....	5
Mean Data.....	6
Construct-related Validity.....	6
Subgroup Differences.....	8
Response Distortion.....	10
Present Study.....	11
Construct-related Validity: Alternative Explanation of Glaze et al.'s (2011) Results.....	12
Hypothesis 1.....	13
Hypothesis 2.....	14
Hypothesis 3.....	14
Subgroup Differences.....	15
Hypothesis 4a.....	15
Hypothesis 4b.....	15
Hypothesis 5a.....	16
Hypothesis 5b.....	16
Response Distortion.....	16
Hypothesis 6a.....	17
Hypothesis 6b.....	17
Test-taker Reactions.....	17
Hypothesis 7a.....	18
Hypothesis 7b.....	18

	Page
Internal Psychometric Properties of the Three	
SJT Response Formats.....	18
Research Question.....	19
METHOD.....	20
Participants.....	20
Measures.....	21
Integrity-based Situational Judgment Test (SJT).....	21
SJT Completion Time.....	23
General Mental Ability (GMA).....	23
Personality.....	26
Response Distortion.....	28
Test-taker Reactions.....	29
Design and Procedure.....	29
Time 1.....	30
Time 2.....	32
RESULTS.....	34
DISCUSSION AND CONCLUSIONS.....	47
Scientific and Practical Implications.....	51
Limitations and Future Directions.....	52
Conclusions.....	55
REFERENCES.....	56
APPENDIX A: TEST BATTERY SCALE SAMPLE ITEMS.....	66
APPENDIX B: PILOT TESTING FOR THE THREE RAVEN’S ADVANCED	
PROGRESSIVE MATRICES RESPONSE FORMATS.....	74
APPENDIX C: SCORING KEY FOR THE NORMATIVE DATA SCORING	
METHOD OF THE GMA TEST.....	76
APPENDIX D: ALTERNATE SCORING METHOD FOR THE RESPONSE	
FORMATS.....	78
APPENDIX E: TIME INTERVAL ANALYSIS.....	83

LIST OF TABLES

	Page
Table 1. Integrity-Based Situational Judgment Test Correlations with General Mental Ability and the Specified Five-Factor Model Personality Dimensions for the Three Response Formats.....	36
Table 2. Descriptive Statistics for Variables That Used Their Standard Response Formats.....	38
Table 3. White–Hispanic Subgroup Differences for the Integrity-Based Situational Judgment Test for All Response Formats.....	43
Table 4. Sex-Based Subgroup Differences for the Integrity-Based Situational Judgment Test for All Response Formats.....	43
Table 5. Descriptive Statistics for the Three Response Formats.....	44
Table 6. Internal Consistency, Test-Retest, and Alternate-Form Reliabilities for the Integrity-Based Situational Judgment Test Scores.....	46
Table C.1. Descriptive Statistics for the SME and Normative Data Scoring Keys of the GMA Test.....	77
Table D.1. Integrity-Based Situational Judgment Test Correlations with General Mental Ability for the Alternate Scoring Method of the Three Response Formats.....	79
Table D.2. Descriptive Statistics for the Absolute and Partial Credit Scoring Methods of the Integrity-Based Situational Judgment Test and GMA Test.....	80
Table D.3. White–Hispanic Subgroup Differences Using the Alternate Scoring Method for the Integrity-Based Situational Judgment Test for All Response Formats.....	81
Table D.4. Sex-Based Subgroup Differences Using the Alternate Scoring Method for the Integrity-Based Situational Judgment Test for All Response Formats.....	81

	Page
Table D.5. Internal Consistency, Test-Retest, and Alternate-Form Reliabilities for the Alternate Scoring Method of the Integrity-Based Situational Judgment Test Scores.....	82
Table D.6. Alternate-Form Reliabilities for the Alternate Scoring Method of the GMA Test.....	82
Table E.1. Retest Interval Length Frequencies.....	83

LIST OF FIGURES

	Page
Figure 1. Study Research Design and Participant Assignments.....	30
Figure 2. Time 1 Score Distributions for the Three SJT Response Formats.....	45
Figure 3. Time 2 Score Distributions for the Three SJT Response Formats.....	45

INTRODUCTION AND LITERATURE REVIEW

The use of situational judgment tests (SJTs) is increasingly common in personnel selection and other organizational decision-making contexts. SJTs are low-fidelity simulations in which test-takers are presented with work-related situations and a set of predetermined responses (Motowidlo, Dunnette, & Carter, 1990). In completing an SJT, respondents report how they would or should handle the situation presented in each item stem, using the predetermined response options (Motowidlo et al., 1990; Ployhart & MacKenzie, 2011). A glaring gap in the extant literature is the absence of any guidance as to which SJT response format to use in practice and under what conditions. Consequently, the objective of the present study was to investigate the construct-related validity of three SJT response formats as a constructive replication and extension of the findings of Glaze, Jarrett, Arthur, Schurig, and Taylor (2011).

Strengths of Situational Judgment Tests

A number of advantages accompany the use of SJTs in personnel selection and organizational decision-making. It should be noted that the summaries presented below are somewhat conceptually problematic because they typify the classic construct/method confound (Arthur & Villado, 2008), such that it is impossible to ascertain whether the described effects are method- or construct-effects. That said, the extant literature indicates that SJTs display moderately strong criterion-related validities in predicting job performance (McDaniel, Morgeson, Finnegan, Campion, & Braverman, 2001; Ployhart & Holtz, 2008). Second, the literature suggests that SJTs demonstrate greater face

validity (which is related to increased motivation for test-taking; Bauer & Truxillo, 2006), and may be capable of engendering smaller race-based subgroup differences than tests of general mental ability (GMA; McDaniel & Nguyen, 2001; Ployhart & Holtz, 2008; Ployhart & MacKenzie, 2011; Whetzel, McDaniel, & Nguyen, 2008; cf. Arthur, Doverspike, Barrett, & Miguel, 2013). Third, evidence from previous research shows that SJTs demonstrate incremental validity in predicting job performance over cognitive ability, conscientiousness and other Big Five personality traits, job experience, and job knowledge (Chan & Schmitt, 2002; Clevenger, Pereira, Wiechmann, Schmitt, & Harvey, 2001). However, as previously noted, what remains unresolved or unestablished is whether these observed effects are method- or construct-level effects.

Situational Judgment Tests in Practice

As low-fidelity simulations, SJTs are relatively inexpensive to develop, administer, and score (Clevenger et al., 2001) when compared to high-fidelity simulations (e.g., assessment centers). However, when additional design features are added to increase their fidelity (e.g., video/interactive SJT formats) their cost correspondingly increases. Furthermore, a number of issues should be considered prior to SJT development and implementation. First, SJT items are often context-bound (Ployhart & MacKenzie, 2011). Hence, many items may be job-, organization-, or industry-specific and thus, may not be generalizable across work situations. Second, there is the potential for the presence of multiple and competing goals in item responding that must be balanced or prioritized (Ployhart & MacKenzie, 2011). Specifically, response options within an item may conflict between employee and

organizational objectives. In addition, test-takers may be inclined to engage in response distortion by answering in a manner that they consider to be aligned with what the organization wants or is interested in. Third, test developers must address issues related to item development (Ployhart & MacKenzie, 2011). In particular, generating a job-relevant hypothetical situation for the item stem, and originating and selecting appropriate response options are crucial to content-related validity. Relevant to this, test developers should exhibit caution when determining both test length (Cortina, 1993) and response ordering within items, because construct-irrelevant order effects can increase test difficulty (Marentette, Meyers, Hurtz, & Kuang, 2012). Scoring the response options should also be managed cautiously to ensure the psychometric validity of the test.

Methods vs. Constructs

It is important to clarify that SJTs are a measurement method (Motowidlo et al., 1990). Relatedly, the distinction between methods and constructs, when comparing predictors in personnel selection, is an important one (Arthur & Villado, 2008; Schmitt & Chan, 2006). The SJT is a predictor method that can be designed to measure a predetermined individual construct, or number of constructs (e.g., job knowledge, interpersonal skills, teamwork, leadership, conscientiousness, agreeableness, and/or emotional stability; Ployhart & MacKenzie, 2011). Hence, SJT development should take place only after one identifies the construct(s) to be assessed.

Several design characteristics and features influence SJT efficacy (Arthur & Villado, 2008; Christian, Edwards, & Bradley, 2010; Glaze et al., 2011; Truxillo, Seitz,

& Bauer, 2008). First, previous research has discussed the impact of the mode of presentation (i.e., written, verbal, video-based, computer-based) and level of fidelity on SJT effectiveness (Chan & Schmitt, 1997; Chan & Schmitt, 2002; Clevenger et al., 2001; Olsen-Buchanan, Drasgow, Moberg, Mead, Keenan, & Donovan, 1998; Weekley & Jones, 1997). Second, response instructions (e.g., behavioral- [“would do”] versus knowledge-based [“should do”]) have been found to influence the construct-related validity of SJTs (McDaniel, Whetzel, Hartman, Nguyen, & Grubb, 2006; Nguyen, Biderman, & McDaniel, 2005; Ployhart & Ehrhart, 2003). Specifically, Ployhart and Ehrhart (2003) found that behavioral-based instructions displayed a strong relationship with personality scores but not with GMA, while the inverse was found for knowledge-based instructions. However, the construct measured by the SJT in their study was not reported, which complicates the interpretation of these findings. Furthermore, their effects are quite dramatic, given the subtle manipulation in the response instructions, so a replication of these results is warranted. Third, the scoring strategies utilized have implications for the predictive validity of these tests (Bergman, Drasgow, Donovan, Henning, & Juraska, 2006; Cullen, Sackett, & Lievens, 2006). Fourth, item stem complexity can produce differential responding between individuals (McDaniel & Nguyen, 2001; Ployhart & MacKenzie, 2011). Specifically, the more complex the item stem, the more difficulty test-takers may have in discerning the item content and providing an informed response.

Although these characteristics of SJTs have been examined in previous research, one design feature that has received little attention in the extant literature is the item

response format (Glaze et al., 2011). Response format refers to the manner in which test-takers respond to test items. The three most commonly used response formats in the extant literature are the rate, rank, and most/least formats. For the *rate response format*, respondents rate (typically using a Likert scale) the effectiveness of each response option as a solution to the situation presented in an item stem. The *rank response format* requires respondents to rank order the effectiveness of all response options within each item, and the *most/least response format* involves selecting the most effective and least effective response options within each item. Glaze et al. argued that the rank format engenders the most cognitive load, followed by the most/least, and then the rate.

Initial Research on Situational Judgment Test Response Formats (Glaze, Jarrett, Arthur, Schurig, & Taylor, 2011)

One result of the limited attention to SJT response formats is that the extant literature does not offer any guidance as to which response format to use under particular conditions or situations. Thus, Glaze et al. (2011) were first to provide this guidance. Their study assessed an applicant sample ($N = 31,194$), utilizing a between-subjects design (rate-SJT $n = 10,421$; rank-SJT $n = 10,345$; most/least-SJT $n = 10,388$). Applicants completed a 20-item integrity-based SJT using one of three response formats (i.e., rate, rank, and most/least), as well as a GMA test and a personality measure in an unproctored internet-based selection battery assessment. The authors found high levels of internal consistency for all three SJT response format scores, with reliability estimates of .95, .87, and .91 for the rate-, rank-, and most/least-SJTs, respectively.

Mean data. Glaze et al. (2011) obtained a number of important results.

Primarily, they observed different mean SJT scores for the three response formats.

Specifically, the mean SJT score for the most/least format ($M = 82.85$, $SD = 11.96$) was greater than that of the rate format ($M = 60.06$, $SD = 16.87$, $d = 1.96$, $p < .05$), which in turn was greater than the rank format ($M = 53.19$, $SD = 15.44$, $d = 0.42$, $p < .05$). This finding indicates that it is easier to identify the most and least effective responses to an SJT scenario than to either rate or rank the responses, with the rank format being the most difficult. Furthermore, the rank-SJT displayed better distributional properties than the rate- and most/least-SJTs, providing a better approximation of a normal distribution of scores. In summary, it may be easier to identify correct responses using a most/least-SJT than a rate-SJT, which in turn may be easier than the rank-SJT. However, the most/least-SJT displayed a rather large mean and the smallest standard deviation among the three response formats, and the data were also very negatively skewed, suggesting the possibility of range restriction with a lower variability in scores for this response format.

Construct-related validity. Glaze et al.'s (2011) results also indicated that the relationship between GMA and integrity-based SJT scores varied as a function of the response formats. The rank-SJT displayed a stronger correlation with GMA (.28) than did the most/least-SJT (.25, $z_r = 2.16$, $p < .05$), which in turn displayed a stronger correlation with GMA than did the rate-SJT (.16, $z_r = 6.48$, $p < .05$), supporting their hypothesized effects. However, integrity is a noncognitive construct, and thus should not covary highly with GMA ($\rho = .02$; Ones, 1993), so the preferred outcome is one

where integrity-based SJT scores display a weak, or ideally, zero relationship with GMA. Thus, Glaze et al. suggested that the higher cognitive and information-processing demands associated with the rank and most/least response formats are reflected in their stronger relationship with GMA.

In contrast, integrity tests generally display meaningful relationships with the Big Five factors of agreeableness, conscientiousness, and emotional stability (Berry, Sackett, & Wiemann, 2007). Consequently, Glaze et al. (2011) investigated the relationship between the response formats of the integrity-based SJT and these personality traits. In support of their hypothesis, the results showed that the rate-SJT correlated with the specified personality traits more strongly than did the most/least-SJT. Specifically, the rate-SJT–agreeableness correlation was greater than the most/least-SJT–agreeableness correlation ($z_r = 7.09, p < .05$), the rate-SJT–conscientiousness correlation was greater than the most/least-SJT–conscientiousness correlation ($z_r = 8.78, p < .05$), and the rate-SJT–emotional stability correlation was greater than the most/least-SJT–emotional stability correlation ($z_r = 7.40, p < .05$). Smaller effects were found for the most/least- and rank-SJTs in their correlations with the specified personality traits. The most/least-SJT displayed slightly stronger correlations than the rank-SJT with agreeableness ($z_r = 1.50, p > .05$), conscientiousness ($z_r = 0.75, p > .05$), and emotional stability ($z_r = 0.73, p > .05$), but because none of the most/least–rank comparisons were significant, their hypothesis that the most/least-SJT would correlate more strongly than the rank-SJT with the specified personality traits was not supported. It should be noted that the magnitudes of the rate-SJT relationships with the specified

personality traits were similar to those observed for traditional integrity tests.

Additionally, the weaker relationships obtained for the rank and most/least response formats may be attributable to the increased cognitive and information-processing demands associated with these formats. In summary, their pattern of results indicated that, consonant with the construct measured by the SJT (i.e., integrity), scores on the rate response format were more related to an individual's agreeableness, conscientiousness, and emotional stability scores than the most/least and rank response formats.

Additional support for the differential cognitive and information-processing demand premise posited by Glaze et al. (2011) was provided by the differences in completion times. Specifically, the rank-SJT took participants longer to complete ($M = 20.83$ minutes, $SD = 9.09$) than the most/least-SJT ($M = 17.76$, $SD = 8.10$; $d = 0.36$, $p < .05$), which in turn took longer to complete than the rate-SJT ($M = 14.35$, $SD = 7.18$; $d = 0.45$, $p < .05$).

Subgroup differences. An important issue of interest to both researchers and practitioners is the extent to which selection tests display subgroup differences in test scores, which often translate into adverse impact. Race- and sex-based differences are among the most widely investigated classes of subgroup differences in personnel selection and testing. Glaze et al. (2011) investigated race-, sex-, and age-based subgroup differences. However, because of the inability to obtain a race- and age-diverse sample, the present study did not investigate age-based subgroup differences, and race-based differences were limited to Hispanic/White comparisons. Sex-based differences were investigated.

In their investigation of race-based subgroup differences, Glaze et al. (2011) found greater differences for the rank and most/least response formats than for the rate response format. Of note, the rank-SJT displayed the largest subgroup difference when comparing Hispanic applicants to White applicants ($d = 0.47$), followed by the most/least-SJT ($d = 0.43$), and then the rate-SJT ($d = 0.20$). All d s were statistically significant ($p < .05$), with positive d s indicating that Whites scored higher than Hispanics. These findings suggest that the response format can impact the magnitude of race-based subgroup differences in SJT scores, and that regardless of the response format, these differences tend to favor White respondents.

Past research has found that females tend to score higher on integrity tests, compared to males (d s range from 0.11 to 0.27; Berry et al., 2007). In addition, although the specific constructs were unspecified, Whetzel, McDaniel, and Nguyen (2008) reported that females had higher SJT scores than males. Consistent with the extant literature, Glaze et al.'s (2011) results indicated that women obtained higher scores than men on all three SJT response formats. Furthermore, their results demonstrated that the rank-SJT resulted in a sex-based subgroup difference ($d = -0.21$) that was larger than that for the most/least-SJT ($d = -0.09$, $z = 194.89$, $p < .05$), and the rate-SJT ($d = -0.08$, $z = 197.88$, $p < .05$). The difference between the rate- and the most/least-SJT was also significant ($z = 16.14$, $p < .05$).

In summary, it appears that the SJT response formats engender varied degrees of race- and sex-based subgroup differences that tend to favor White and female

respondents, respectively. Furthermore, increases in cognitive ability loading for SJTs correspondingly widen these gaps.

Response distortion. Response distortion is defined as the tendency for an individual to intentionally over-report socially desirable personal characteristics and to intentionally under-report socially undesirable characteristics (Paulhus, 2002). Simply put, response distortion is the deliberate tailoring of answers to create a positive impression (Paulhus, 1991a; Rosse, Stecher, Miller, & Levin, 1998; Zerbe & Paulhus, 1987). Glaze et al.'s (2011) response distortion data displayed moderately negative correlations with SJT scores. In general, individuals with higher response distortion scores had lower levels of integrity, as reflected in their SJT scores. The relationship between response distortion and the rate-SJT scores ($-.22$) was significantly different from that for the most/least-SJT scores ($-.33$; $z_r = 7.92$, $p < .05$), supporting their hypothesis that the rate-SJT would be more susceptible to response distortion than the most/least-SJT. However, the relationship between response distortion and the most/least-SJT was not significantly different from that for the rank-SJT ($-.34$; $z_r = 0.72$, $p > .05$), so their hypothesis that the most/least-SJT would be more susceptible to response distortion than the rank-SJT was not supported. These findings suggest that if test-takers complete an SJT with a rate response format, they are more likely to exhibit higher levels of response distortion than if they complete an SJT with a most/least or rank response format because the latter two are similarly more resilient to engendering response distortion. Thus, as noted by Glaze et al., if response distortion is a major concern of test developers, the rate response format may not be the desired format to use.

It should be noted that Glaze et al.'s (2011) results are at odds with what would be expected of a response distortion measure. Specifically, response distortion scores should be positively related to integrity, such that high response distortion should inflate SJT scores because respondents seek to answer items in a manner that will leave a positive impression. In fact, a potential critique of Glaze et al.'s response distortion measure is that its results are consonant with the extant response distortion literature only if it is assumed to be an alternative measure of integrity, rather than that of response distortion. Furthermore, the measure used an inconsistency-in-responding approach to operationalize response distortion. However, there is some question as to whether inconsistency-in-responding is best conceptualized as a response set (i.e., as a means of detecting response distortion) or as a response style (i.e., as a means of detecting careless responding). Consequently, the present study sought to replicate the findings of Glaze et al., and to clarify their findings using a more traditional measure of response distortion.

Present Study

Despite the findings of Glaze et al. (2011), some characteristics of their study design present a number of issues that warrant further investigation. Therefore, the present study sought to undertake a constructive replication and extension of Glaze et al. to address a potential common method bias threat to their findings, and in so doing, investigate a number of additional issues. Specifically, a “shared-common-response-method effect” serves as a plausible alternative explanation for the observed effects instead of the posited differences-in-g-loading explanation, which suggests that the variance in scores is due to the differences in the cognitive and information-processing

demands of each format. Thus, the present study utilized a design in which an integrity-based SJT using rate-, rank-, and most/least-response formats was crossed with a GMA test and a five-factor model (FFM) personality measure using the same response formats. In addition, the present study extended Glaze et al.'s response distortion results by investigating the comparative susceptibility of the three response formats to response distortion with a different, more commonly used operationalization of response distortion. Sex-based subgroup differences, White–Hispanic subgroup differences, and test-taker reactions, as well as the internal consistency, test-retest, and alternate-form reliabilities of the three response formats were also assessed.

Construct-Related Validity: Alternative Explanation of Glaze et al.'s (2011) Results

Glaze et al. (2011) posited that differences in the cognitive and information-processing demands between response formats caused their observed effects. However, as noted above, the main critique of their findings pertains to a shared-common-response-method effect methodological artifact. The shared-common-response-method effect is a measurement bias in which at least part of the variance accounted for in the relationship between two or more tests can be attributed to the format similarities among the tests (Podsakoff, MacKenzie, Lee, & Podsakoff, 2003). In this case, the rate-SJT shared a similar response format with the personality measure, and the rank- and most/least-SJTs appeared to share a similar response format with the GMA test (i.e., multiple-choice). Consequently, the shared-common-response-method effect serves as a plausible alternative explanation for their observed effects. Nevertheless, Glaze et al.'s subgroup differences data supported their proposition that the response formats produce

varying levels of cognitive and information-processing demands. As they noted, the shared-common-response-method effect explanation would not predict nor account for these observed subgroup differences or their completion time results.

Glaze et al. (2011) present a compelling case for the substantive basis of their findings. Nevertheless, a laboratory-based experiment that simulates conditions in which a personality measure has a response format similar to a GMA [or knowledge] test (i.e., multiple-choice) and a GMA test has a response format similar to a personality measure (i.e., Likert response scale), and investigates their resultant effects, would be informative and provide more conclusive statements about the obtained results. Therefore, the present study investigated the construct-related validity of the three SJT response formats as a constructive replication and extension of the findings of Glaze et al. In summary, the primary objective of the present study was to provide a comparative test of the differences-in-g-loading versus shared-common-response-method explanations of Glaze et al.'s observed effects. Hence, it was hypothesized that:

Hypothesis 1. *If the differences-in-g-loading explanation best accounts for the observed response format effects, then the rank-SJT will demonstrate stronger correlations with GMA than the most/least-SJT, which will in turn display stronger correlations with GMA than the rate-SJT (regardless of the GMA test response format). However, if the results are due to the shared-common-response-method effect, then these effects will be obtained for only the conditions in which the SJT response format matches that of the GMA test response format (especially in the rank condition).*

Hypothesis 2. *If the differences-in-g-loading explanation best accounts for the observed response format effects, then the rate-SJT will demonstrate stronger correlations with agreeableness, conscientiousness, and emotional stability than the most/least-SJT, which in turn will show stronger correlations than the rank-SJT (regardless of the personality measure response format). However, if the results are due to the shared-common-response-method explanation, then the agreeableness, conscientiousness, and emotional stability effects will be obtained for only the conditions in which the SJT response format matches the personality measure response format (especially in the rate condition).*

To further investigate the construct-related validity of the SJT, the response formats were compared in terms of completion times. Consonant with Glaze et al.'s (2011) results, one would expect different completion times for the SJT response formats; namely that ranking an item's response options should require more time to complete than selecting the most and least effective response options, which in turn should require more time to complete than rating each response option. This is consistent with the proposition that test items with greater cognitive demands elicit longer response latency (Basilli & Scott, 1996; Yan & Tourangeau, 2008). Therefore, it was hypothesized that:

Hypothesis 3. *The rank-SJT will demonstrate the longest completion time, followed by the most/least-SJT, and then the rate-SJT.*

Subgroup Differences

Because integrity tests show negligible race-based subgroup differences (Ones, 1993), it is reasonable to expect similar differences on an integrity-based SJT. However, it is also recognized that cognitively loaded tests display race-based subgroup differences (Roth, Bevier, Bobko, Switzer, & Tyler, 2001). Specifically, in the context of the comparisons undertaken in the present study, Hispanics tend to score lower on cognitively loaded tests than Whites. Whetzel et al. (2008) found that SJTs that load higher on cognitive ability produced larger White–Hispanic differences. In addition, Ployhart and Holtz (2008) reported that SJTs (constructs unspecified) show subgroup differences favoring Whites over Hispanics. Consistent with the extant literature, Glaze et al.'s (2011) results indicated differences in SJT scores between Whites and Hispanics. Furthermore, the subgroup difference for the rank-SJT was larger than that of the most/least-SJT, which in turn was larger than the rate-SJT. Consequently, in accordance with the results of Glaze et al., it was hypothesized that:

Hypothesis 4a. *The rank-SJT will display larger White–Hispanic differences compared to the most/least-SJT.*

Hypothesis 4b. *The most/least-SJT will display larger White–Hispanic differences than the rate-SJT.*

The extant literature has demonstrated that in the context of sex-based subgroup differences, females tend to outperform males on integrity tests (Berry et al., 2007) and SJTs (Whetzel et al., 2008). Consonant with this, Glaze et al. (2011) found that women displayed higher scores than men on all three integrity-based SJT response formats, and

that the rank-SJT resulted in a larger sex-based subgroup difference than the most/least-, and then the rate-SJT. Thus, it was hypothesized that:

Hypothesis 5a. *the rank-SJT will display larger sex-based differences than the most/least-SJT.*

Hypothesis 5b. *The most/least-SJT will display larger sex-based differences compared to the rate-SJT.*

Response Distortion

Another objective of the present study was to explore the effect of response distortion on the SJT response formats, using a less controversial and more widely accepted measure of response distortion. As previously noted, Glaze et al.'s (2011) operationalization of response distortion is potentially problematic. Using the personality measure, a response distortion score was generated using 10 personality test item pairs (one positive and one negative) to assess the consistency in test-takers' responses. This method of detecting response distortion is similar to that of the Guilford-Zimmerman Temperament Survey (Guilford, Zimmerman, & Guilford, 1976). After reverse coding, the response distortion score was computed as the absolute difference in responses on the pairs of consistent items such that larger scores reflect higher levels of response distortion. Validity evidence of this approach to measuring response distortion has been demonstrated in previous research. For example, Hand's (1964) social desirability scale, which is comprised of 20 item pairs from the Guilford-Zimmerman Temperament Survey, demonstrated a meaningful relationship ($r = .54$) with the Edwards Social Desirability Scale (Edwards, 1957).

Although this method is often employed to detect response distortion, it is also used to examine consistency-in-responding as a means to assess whether respondents were paying attention to the items, and to identify random responding. This approach is present in several scales, including those developed for normal populations (e.g., NEO-PI-R; Kurtz & Parrish, 2001) and clinical populations [e.g., Variable Response Inconsistency (VRIN) and True Response Inconsistency (TRIN) scales of the MMPI; Berry et al., 1992; Wetter, Baer, Berry, Smith, & Larsen, 1992]. Thus, as previously noted, there is a reasonable debate as to whether inconsistency-in-responding is best conceptualized as a response style or as a response set (e.g., Bond, 1986; 1987). Consequently, to address the ambiguity with the interpretation of Glaze et al.'s (2011) measure, the present study investigated the comparative susceptibility of the response formats to response distortion by using a less debatable and more widely accepted measure of response distortion. However, the expectation was to find a pattern of results for response distortion similar to Glaze et al.'s conclusions, such that:

Hypothesis 6a. *The rate-SJT will display stronger correlations with response distortion scale scores compared to the most/least-SJT.*

Hypothesis 6b. *The most/least-SJT will display stronger correlations with response distortion scale scores compared to the rank-SJT.*

Test-Taker Reactions

It is useful to consider test-taker reactions to selection methods as an important comparative evaluation criterion because they are associated with outcomes such as perceptions of fairness, face validity, test-taking motivation, test performance, and self-

withdrawal from the selection process (Anderson, 2004; Arthur & Villado, 2008; Chan & Schmitt, 2004; Ryan & Ployhart, 2000). Increasing a test's complexity correspondingly increases its cognitive load, which is likely to exhibit an inverse relationship with respondents' reactions to the test. Specifically, it was posited that respondents would react more favorably to the response formats with less cognitive load. Prior work has shown that differing item formats can produce differences in test-taker reactions (Shyamsunder & McCune, 2009). Furthermore, Bradshaw (1990) found that the perceived difficulty in processing of a placement test affected individuals' reactions to the test, such that test-takers reacted more negatively to the placement test as the perceived difficulty in processing of the test correspondingly increased. As previously described, the information-processing demands of the most/least- and particularly the rank-SJT increase the cognitive load of the test formats above that of the rate-SJT. Thus, in comparing the SJT response formats in terms of test-taker reactions, it was posited that:

Hypothesis 7a. *Participants will react more positively to the rate-SJT than the most/least-SJT.*

Hypothesis 7b. *Participants will react more positively to the most/least-SJT than the rank-SJT.*

Internal Psychometric Properties of the Three SJT Response Formats

Additionally, the internal psychometric properties of the three SJT response formats were assessed. First, descriptive statistics for the SJT response formats were obtained. Next, the internal consistency, test-retest, and alternate-form reliability

estimates of the three response format scores for the SJT were obtained. Specifically, the consistency of respondents' scores across the same and different SJT response formats over time was assessed to investigate both the temporal and alternate-form stability of the observed test scores as a function of the response formats. Because differences in response format can impact SJT scores, the potential exists for the psychometric properties to also differ between the response formats. For instance, Glaze et al. (2011) looked at the internal consistency reliabilities and the obtained differences between the three SJT response format scores. However, the nature of the present study sample differed from theirs in such a way that an additional investigation of the internal psychometric properties of the SJT response formats would be an informative addition to the extant SJT literature. Therefore, the following research question was posed:

Research Question. *Do the internal psychometric properties of the integrity-based SJT differ by response format?*

METHOD

Participants

An initial sample of 505 undergraduate students were recruited from the Texas A&M University Subject Pool to attend two proctored research sessions and complete a battery of paper-and-pencil tests during each session. However, a number of test-taker errors were later revealed, specifically that many participants did not follow the written and verbal instructions for correctly completing the GMA test response formats. Also, 23 participants did not return to complete the second study session, resulting in incomplete data for these cases. Therefore, the participants with errors and/or incomplete data were removed from the sample and an additional 137 participants were recruited to meet the predetermined sample size as per the a priori power analysis. To minimize test-taker error, more thorough verbal instructions for the GMA test were provided to these participants during their study sessions, and as a result, all of the new cases were retained with complete data. Therefore, of the total 642 participants recruited, complete and usable data were obtained for 492, who were retained as the final study sample. Two-thirds of the final sample were female ($n = 327$, 66.5%), and the mean age was 18.77 ($SD = 1.47$). The final sample consisted of 330 (67.1%) Caucasians, 18 (3.7%) African-Americans, 86 (17.2%) Hispanics, 35 (7.1%) Asian-Americans, 5 (1.0%) individuals of other ethnic groups, and 18 (3.6%) multiracial individuals. Students participated in exchange for course-related research credit.

Measures

Integrity-based situational judgment test (SJT). The present study used the same 20-item loss prevention test used by Glaze et al. (2011). This test was designed to assess an individual's propensity to engage in counterproductive work behaviors (CWBs) that would result in monetary loss to the organization. For this SJT, there were three different response formats. The rate response format involved participants rating the effectiveness of each of five response options for each item on a five-point, Likert scale (1 = highly ineffective action or response, 5 = highly effective action or response). For the rank response format, participants ranked the effectiveness of the five response options for each item from 1 (most effective) to 5 (least effective). Finally, the most/least response format called for participants to identify the most effective and least effective actions from the five available response options for each item by placing an "M" next to the most effective action and an "L" next to the least effective action. The SJT response formats displayed internal consistency reliability estimates of .94, .76, and .69 at Time 1, and .95, .81, and .76 at Time 2 for the rate-, rank-, and most/least-SJTs, respectively. Sample items for the SJT response formats are found in Appendix A.

The SJT scoring keys used by Glaze et al. (2011) were generated by a panel of six industry professionals identified as subject matter experts (SMEs). For the rate-SJT, the SMEs independently rated each response option in terms of its effectiveness on a 5-point scale (1 = very ineffective; 5 = very effective). Keyed responses were the aggregate of the SME ratings, rounded to the whole number. One point was awarded for

each response rating that matched the keyed SME rating, so test-takers received between 0 and 5 points for each test item, and test scores could range from 0 to 100 points.

For the rank-SJT, the SMEs independently ranked the response options from the most effective to the least effective. The average SME ranking within each test item was used to designate the keyed rank order. One point was awarded for each response ranking that matched the SME ranking, so test-takers received between 0 and 5 points for each item, and test scores could range from 0 to 100.

For the most/least-SJT, the highest and lowest ranked response options from the rank-SJT scoring key were designated as the most and least effective response options, respectively. The scoring algorithm developed by Motowidlo, Dunnette, and Carter (1990) was used for this response format. Specifically, test-takers were awarded one point each for correctly identifying the most and least effective response options. Test-takers lost one point each for identifying the most effective response option as the least effective, and the least effective response option as the most effective. No points were awarded or taken away for identifying a non-keyed response option as either most or least effective. Consequently, test-takers' scores could range from -2 to 2 for each item. Item scores were summed across the 20 scenarios such that test scores could range from -40 to 40. For ease of interpretation and comparative purposes, test scores were scaled to 100 to match the score range of the rate- and rank-SJT scoring keys.

For the sake of completeness, an alternative scoring method for the SJT to the one used in the primary analyses in which partial credit was given for the rate and rank

response options was considered, and the results of the comparison between the scoring methods are presented in Appendix D.

SJT completion time. Completion times for the SJTs were obtained via participant self-report. Specifically, participants were instructed to record their start and end times in their test booklets. The process and procedures for obtaining these data are described in detail in the *Procedure* section.

General mental ability (GMA). GMA was operationalized as scores on the short form of the Raven's Advanced Progressive Matrices (APM; Arthur & Day, 1994; Arthur, Tubre, Paul, & Sanchez-Ku, 1999) which consists of 2 practice items and 12 test items. This is regarded as a test with a low level of culture loading, and it is considered the purest available test of fluid intelligence (Raven, 1989, 2000). Arthur et al. (1999) reported a 1-week test-retest reliability of .76 for this test.

Three different response formats of the APM were developed for the present study, and sample items are presented in Appendix A. The rate response format required participants to rate how accurately each of the eight answer options fit the item pattern to correctly answer the item using a five-point Likert scale (1 = very poor fit, 5 = very good fit). Participants completing the rank response format ranked the eight response options (1 = best fit, 8 = worst fit) on how accurately they fit the item pattern to correctly answer the question. For the most/least response format, participants indicated the response option that most accurately completed the item pattern and the response option that least accurately completed the item pattern, placing a "B" next to the response option with the best fit and a "W" next to the response option with the worst fit. On the basis of two

rounds of pilot testing (see Appendix B), the rate, most/least, and rank response formats of the APM were administered at Time 1 of data collection with a 30 minute time limit. The standard short-form APM format was administered at Time 2 of data collection with the corresponding 15 minute time limit.

The APM response format scoring keys were developed for the present study, using a panel of seven upper-level industrial/organizational (I/O) psychology graduate students as SMEs. The SMEs generated the keys via consensus (see Appendix C). For the rate response format, the SMEs individually rated all response options within each test item on 5-point scale (5 = very good fit) prior to the consensus meeting. The response option with the best fit (i.e., the keyed correct answer as per the test manual) was rated '5' and all other response options were rated at lower points on the scale. Based on the initial level of agreement among the panel for each response option, discussions took place as necessary, with the purpose of reaching consensus for the appropriate rating of each response option. Agreed upon ratings for all response options within each item were identified as the keyed responses, and the SME consensus ratings produced the scoring key for the rate response format. The same scoring method was used as with the rate-SJT. Specifically, one point was awarded for each response rating that matched the keyed SME rating, so test-takers received between 0 and 8 points for each test item, and test scores could range from 0 to 96 points. For ease of interpretation and comparative purposes, test scores were scaled to 100 to match the score range of the SJT response formats.

For the rank response format of the APM, the SMEs used the keyed ratings from the rate response format to rank order the response options from best to worst fit (1 = best fit). The response option rated '5' (i.e., the keyed correct answer as per the test manual) for each item was ranked '1' and all other response options were ranked below that, based on their fit ratings. In the event of ties in ratings between two or more response options within an item, SME discussions led to agreement on the appropriate rank order of the response options in question. The consensus rankings for each item produced the scoring key for the rank response format. The same scoring method was used as with the rank-SJT. Specifically, one point was awarded for each response ranking that matched the SME ranking, so test-takers received between 0 and 8 points for each item, and test scores could range from 0 to 96. For ease of interpretation and comparative purposes, test scores were scaled to 100 to match the score range of the SJT response formats.

For the most/least response format of the APM, the highest and lowest ranked response options from the rank-GMA scoring key were designated as the most and least effective response options, respectively. The same scoring method was used as with the most/least-SJT. Specifically, test-takers were awarded one point each for correctly identifying the most and least effective response options. Test-takers lost one point each for identifying the most effective response option as the least effective, and the least effective response option as the most effective. No points were awarded or taken away for identifying a non-keyed response option as either most or least effective. Consequently, scores could range from -2 to 2 for each item. Item scores were summed

across the 12 items such that test scores could range from -24 to 24. For ease of interpretation and comparative purposes, test scores were scaled to 100 to match the score range of the SJT response formats.

For the sake of completeness, an alternative approach to scoring the APM in which partial credit was given for the rate and rank methods was considered. Hence, a comparative analysis of the method used in the study against the alternative method was conducted, and the results are presented in Appendix D.

Personality. The present study used a 50-item Five Factor Model (FFM) International Personality Item Pool (IPIP) measure (Goldberg, 1999; Goldberg et al., 2006). Whereas the whole measure was administered to participants, only the agreeableness, conscientiousness, and emotional stability factors were examined for the purposes of the present study. Goldberg (1992; 1999) reported internal consistency reliability estimates of .82, .79, and .86 for agreeableness, conscientiousness, and emotional stability scores, respectively. Pertaining to the FFM response formats, the present study obtained high internal consistency reliability estimates of .80, .77, and .74 for agreeableness scores, .80, .72, and .78 for conscientiousness scores, and .83, .84, and .78 for emotional stability scores for the rate-, rank-, and most/least-FFMs, respectively. Although the administration of the personality measures was not timed, participants were asked to record their start and end times in their test booklets.

Along with the standard Likert response format, two additional response formats were developed for the FFM measure. A sample item for each response format can be found in Appendix A. The rate (i.e., standard) response format for the IPIP asked

participants to rate each item on a five-point Likert scale (1 = very inaccurate, 5 = very accurate) in terms of the extent to which each item statement described the behavior of the test-taker. Items were scored from 1 to 5, with the response ‘very inaccurate’ assigned a value of 1, and the response ‘very accurate’ assigned a value of 5. After reverse scoring the specified items, each FFM score was computed as the average of the responses to the 10 items that comprised the specified dimension. Therefore, scores could range from 1 to 5 for each factor.

For the rank response format of the FFM measure, each item was responded to using a frequency-based response format, in which participants estimated the relative frequency that each of the five response levels (ranging from very inaccurate to very accurate) described their behavior (Edwards & Woehr, 2007). Participants assigned a percentage to each response level accordingly, such that the percentages summed to 100%. Participants were instructed to not assign equal percentages to any two response levels, thus forcing the rank ordering of response levels. Responding to items in this manner required that the percentages assigned to each response level be combined into a single score for each item (Edwards & Woehr, 2007). Therefore, within an item, each response level was assigned a weight (very inaccurate = .01, inaccurate = .02, neither inaccurate nor accurate = .03, accurate = .04, very accurate = .05), and the sum of the weighted percentages provided an item score that ranged from 1 to 5 to match the scoring scale for the rate response format. After reverse scoring the specified items, the score for each dimension was computed as the average of the scores for the 10 items

corresponding to the specified factor (i.e., agreeableness, conscientiousness, and emotional stability). Therefore, scores could range from 1 to 5 for each factor.

For each item in the most/least response format of the FFM measure, participants selected from the five response levels (ranging from very inaccurate to very accurate) the one that was most descriptive of their behavior, as well as the response level least descriptive of their behavior. Thus, participants placed an “M” next to the level on the scale that most accurately described their behavior for the item, and an “L” next to the level on the scale that least accurately described their behavior for the item. The score for each item was operationalized as the difference between the numerical values of the ‘most’ and the ‘least’ responses. Consequently, item scores could range from -4 to 4 for each item. After reverse scoring the specified items, the dimension score for each factor (i.e., agreeableness, conscientiousness, and emotional stability) was computed as the sum of the scores for the items comprising that factor. Therefore, test scores could range from -40 to 40 for each factor. For ease of interpretation and comparative purposes, test scores were scaled to a 1–5 range for each factor to match the score range of the rate and rank response formats.

Response distortion. Response distortion was operationalized as scores on the 20-item impression management subscale of the Balanced Inventory of Desirable Responding, version 6, Form 40A (BIDR; Paulhus, 1991b; see Appendix A for sample items). Participants rated each item on a seven-point Likert scale (1 = not true, 7 = very true). After reverse scoring the specified items, the test-taker’s score was computed as the number of items rated as a ‘6’ or ‘7’ (Paulhus, 1991b), and so scores could range

from 0 to 20. The extant literature reports internal consistency reliability estimates for the impression management subscale that range from .77 to .86 (Konstabel, Aavik, & Allik, 2006; Paulhus, 1994; Stober & Dette, 2002); the present study obtained a slightly lower estimate of .72.

Test-taker reactions. Three items, rated on a five-point Likert scale (1 = strongly disagree, 5 = strongly agree; see Appendix A) were developed to measure the perceived difficulty of the SJT response format that test-takers completed at Time 2. The average of these three items was used to create a composite score that represented the perceived difficulty of the SJT response format that test-takers completed at Time 2. The three difficulty items displayed an internal consistency reliability estimate of .84.

A single item was developed to assess test-takers' preferences for each of the three SJT response formats. Specifically, participants were presented with a single SJT sample item depicting all three response formats. They were then instructed to rate (on a 1 to 5 point scale) the extent to which they would prefer to use each response format if they had to complete the SJT again. Thus, unlike the difficulty ratings where participants rated only the SJT they completed at Time 2, participants provided preference ratings for all three response formats.

Design and Procedure

The study was posted on a Texas A&M University Department of Psychology online recruitment portal, and students self-selected into participation. No other recruitment efforts were undertaken. Figure 1 presents an illustration of the study design and protocol. As shown in Figure 1, the study used a mixed factorial design.

Participants were assessed at two time points (each lasting 1 ½ hours), and were randomly assigned to complete one of the three SJT response formats at each time point. Consequently, at Time 2, approximately one-third of the participants retested on the same version of the SJT, while the remaining two-thirds completed a different version than the one completed at Time 1. This study design allowed for both between-subjects and within-subjects comparisons. There was a retest interval of 5-9 days between Time 1 and Time 2.

		SJT Response Format			
		Rate	Rank	Most/Least	
TIME 1	GMA1/FFM Response Format	Rate	57	54	53
		Rank	55	53	57
		Most/Least	54	54	55
TIME 2	GMA2, RD, Test-Taker Reactions		149	162	181

Figure 1. Study research design and participant assignments. The numbers in the cells represent the number of participants in each condition. GMA1 = the short form of the Raven's Advanced Progressive Matrices using either the rate, rank, or most/least response format. GMA2 = the short form of the Raven's Advanced Progressive Matrices using the standard response format. FFM = personality measure (i.e., International Personality Item Pool) using either the rate, rank, or most/least response format. RD = response distortion measure (i.e., Balanced Inventory of Desirable Responding).

Time 1. The present study utilized a 3 (SJT: rate, versus rank, versus most/least) × 3 (GMA/FFM: rate, versus rank, versus most/least) between-subjects design at Time 1. Groups of approximately 100 participants attended self-selected, proctored study sessions to complete the Time 1 measures. Participants were randomly assigned to complete one of the three versions of the SJT, which differed only by response format (i.e., rate, rank, or most/least). Participants were also randomly assigned to complete

one of three versions of the GMA test and FFM personality measure, which also differed only by response format (i.e., rate, rank, or most/least). To clarify, for example, if a participant completed the rate-GMA, they also completed the rate-FFM.

On arrival for the study session, participants received a single test booklet containing all of the Time 1 measures, as well as an informed consent form. The test booklets reflected one of the nine possible study conditions illustrated in Figure 1. When the study session began, the proctor described the nature of the study, and then asked participants to sign and date their consent forms. Following collection of the consent forms, the proctor instructed the participants to complete a demographic form and then to wait for further instructions. Once all participants had completed the demographic form, the proctor instructed the participants to complete the first instrument in the test booklet, the APM. The proctor provided verbal instructions for the test and advised participants to read the written instructions. Participants were then given three minutes to review the instructions and complete the two practice items. After completing the practice items (i.e., at the end of the 3 minutes), participants commenced the actual test, which had a time limit of 30 minutes. As previously noted, two waves of pilot testing were conducted in order to determine the appropriate time limit for the three response formats of this test. Participants were told to sit quietly and wait for the remaining participants to complete the test, should they finish before the time limit. When the time limit was reached, the proctor directed those still working to stop writing and put their pencils down.

After completing the APM, participants completed the SJT and personality measures, respectively. The proctor provided verbal instructions for completing the remaining measures, and advised the participants to review the written instructions for each. Additionally, the proctor placed a digital clock on an overhead projector screen at the front of the room and instructed participants to record their start and end times for the SJT and personality measures. Specifically, there were spaces at the beginning and end of the SJT, as well as at the end of the personality measure, for participants to self-report their start and end times for the measures. Participants completed the remaining measures without any further breaks in the protocol, and on completing the entire test booklet, returned their booklets to the proctor and received a study credit slip as proof of their participation in the session. Lastly, the proctor completed an error report form, noting any deviations from the protocol or unusual events that occurred during the study session.

Time 2. After a 5-9 day interval, participants returned for their self-selected Time 2 sessions in groups of approximately 100 each. Participants were randomly assigned to retest on one of the three response formats of the SJT, which provided a within-subjects comparison for the participants who retested the same SJT response format, and an alternate-form comparison for those who retested on a different format. All participants retested on the standard version of the APM. In addition, participants completed the response distortion and test-taker reactions measures.

On arrival for the study session, participants again received a single test booklet containing all the Time 2 measures. As illustrated in Figure 1, the test booklets reflected

one of three possible retest conditions. At the commencement of the session, the study proctor administered the standard short-form version of the APM. The proctor provided verbal instructions for the test and advised participants to read the written instructions. Participants were then given three minutes to review the instructions and complete the two practice items. After completing the practice items (i.e., after three minutes), the participants started the actual test, with a time limit of 15 minutes (Arthur & Day, 1994; Arthur et al., 1999). Participants were told to sit quietly and wait for the remaining participants to complete the test, should they finish before the time limit. When the time limit was reached, the proctor directed those still working to stop writing and put their pencils down.

After completing the APM, participants completed the SJT, as well as the response distortion and test-taker reactions measures, respectively. The proctor provided verbal instructions for completing the remaining measures and advised the participants to review the written instructions for each. Additionally, the proctor placed a digital clock on an overhead projector screen at the front of the room and instructed participants to record their start and end times for the SJT. The study session was proctored in a manner identical to that at Time 1, with the exception that upon completing the session, participants were given a debriefing form to explain the purpose of the study.

RESULTS

Prior to data collection, an initial power analysis indicated that a sample size of 544 participants was needed to detect the specified effects with a power level of .80. However, due to a host of administrative and logistical reasons associated with obtaining a large sample, the decision was made to recruit a sample of 450-475 participants, resulting in an associated power level of .70 to .80. However, because the final study sample size was 492, another power analysis was conducted using G*Power 3.1 (Erdfelder, Faul, & Buchner, 1996; Faul, Erdfelder, Buchnar, & Lang, 2009) to determine the power level associated with this particular sample size. Using the test of the difference between two independent correlations as the most conservative test of the hypotheses, an estimated large effect ($q^1 = .70$), coupled with an alpha of .05, the final study sample size ($N = 492$) resulted in a 86% chance of detecting the observed effects of the relationship between the SJT and GMA response formats, and a 85% chance of detecting the observed effects of the relationship between the SJT and FFM response formats. To further explore the power levels attained by the present sample, an additional power analysis was conducted using the same preceding parameter estimates but instead with an estimated medium effect ($q = .30$), which resulted in an 99% chance of detecting the effects of the SJT–GMA relationship, and an 97% chance of detecting the effects of the SJT–FFM relationship. It should also be noted that because the power

¹ The q statistic represents the z test effect size estimate for differences between two independent correlations.

analyses are based on tests of the most conservative hypotheses, less stringent tests, such as tests of differences between means, should result in power levels greater than those obtained here.

The results of the tests of Hypotheses 1 and 2—the competitive tests for the differences-in-g-loading versus the shared-common-response-method explanations—are presented in Table 1. Concerning Hypothesis 1, if the shared-common-response-method explanation best accounts for the observed effects, then the matched response format pairs (i.e., rate-SJT/rate-GMA, rank-SJT/rank-GMA, and most/least-SJT/most/least-GMA) should display stronger positive relationships than the other (mismatched) conditions. On the other hand, if the differences-in-g-loading explanation best explains the results, then GMA should display stronger correlations with the rank-SJT than the most/least- and rate-SJTs, regardless of the GMA response format. The underlined correlations in Table 1 (i.e., those on the diagonal) represent the matched response formats. As indicated, the obtained pattern of results does not support the shared-common-response-method explanation. First, for the standard (i.e., multiple-choice) GMA test, the pattern of results provide a constructive replication of Glaze et al.’s (2011) findings, in that the rank-SJT displayed the strongest relationship with GMA scores, followed by the most/least-, and then the rate-SJT. Second, in general, the rank-SJT displayed the strongest relationships with GMA scores regardless of the GMA response format, with considerably weaker relationships obtained for the most/least- and rate-GMA scores, again, regardless of format. Thus, for Hypothesis 1, the results failed to provide support for the shared-common-response-method explanation, and seemed to

best fit the differences-in-g-loading explanation for the observed relationships between the SJT response format scores and GMA.

Table 1
Integrity-Based Situational Judgment Test Correlations with General Mental Ability and the Specified Five-Factor Model Personality Dimensions for the Three Response Formats

	SJT Response Format					
	Rate		Rank		Most/Least	
	Time 1	Time 2	Time 1	Time 2	Time 1	Time 2
GMA Response Format						
Standard	-.01	.06	.29*	.36*	.11	.21*
Rate	<u>.06</u>	<u>-.14</u>	.36*	.46*	.12	.11
Rank	-.11	-.08	<u>.33*</u>	<u>.35*</u>	.06	.00
Most/Least	.08	.16	-.04	.10	<u>-.04</u>	<u>.15</u>
FFM Rate Format						
Agreeableness	<u>.30*</u>	<u>.30*</u>	.10	-.10	-.10	-.02
Conscientiousness	<u>.00</u>	<u>.17</u>	-.25*	-.12	.09	-.13
Emotional Stability	<u>.27*</u>	<u>-.06</u>	-.15	-.10	.12	.07
FFM Rank Format						
Agreeableness	.04	-.14	<u>.19</u>	<u>.25*</u>	.17	.34*
Conscientiousness	.19	.02	<u>.11</u>	<u>.13</u>	-.01	.12
Emotional Stability	.02	-.13	<u>-.26*</u>	<u>-.15</u>	-.17	-.10
FFM Most/Least Format						
Agreeableness	.07	.27*	.08	.09	<u>.12</u>	<u>-.03</u>
Conscientiousness	.14	.39*	.17	.40*	<u>.08</u>	<u>.09</u>
Emotional Stability	.15	.05	.04	.29*	<u>-.07</u>	<u>.00</u>
Response Distortion	.30*	.27*	.17*	.28*	.16*	.13*

Note. Underlined correlations represent those that would be expected to be the highest positive correlations, relative to the others, if the results are indicative of a shared-common-response-method effect. * $p < .05$ (one-tailed).

Concerning Hypothesis 2, a similar pattern of correlations was found for the SJT response formats and the specified FFM personality variables (i.e., agreeableness, conscientiousness, and emotional stability). Indeed, the results did not suggest a pattern in which the strongest relationships were obtained for the matched response formats (i.e., the underlined correlations on the diagonal in Table 1), compared to the other (mismatched) conditions. However, the FFM results neither provided a clear constructive replication of Glaze et al.'s (2011) findings. Nevertheless, although they were not all statistically significant and their magnitudes were smaller than that obtained by Glaze et al., in terms of their patterns, the rate-SJT generally displayed stronger relationships with the standard FFM scores (average of Time 1 and Time 2 correlations = .30, .09, and .14 for agreeableness, conscientiousness, and emotional stability, respectively) than the rank-SJT (average of Time 1 and Time 2 correlations = .00, -.19, and -.13) and the most/least-SJT (average of Time 1 and Time 2 correlations = -.06, -.02, and .10). Thus, for Hypothesis 2, the results were again more aligned with the differences-in-g-loading explanation than the shared-common-response-method effect explanation.

Mean SJT completion times for both time points can be found in Table 2. For Hypothesis 3, the completion time results provided a constructive replication of Glaze et al.'s (2011) findings. As indicated in Table 2, at both time points, the rank-SJT took participants longer to complete than the most/least-SJT ($t = 3.86$, $d = 0.43$, $p < .05$ at Time 1, $t = 4.40$, $d = 0.48$, $p < .05$ at Time 2), which in turn took participants longer to complete than the rate-SJT ($t = 2.70$, $d = 0.30$, $p < .05$ at Time 1, $t = 0.45$, $d = 0.05$, $p >$

.05 at Time 2). Therefore, Hypothesis 3 was supported, and consonant with the expected retest effect, the completion times for all SJT response formats were shorter at Time 2 than at Time 1.

Table 2
Descriptive Statistics for Variables That Used Their Standard Response Formats

Variable	Mean	SD
GMA	68.31	19.52
Response Distortion	6.85	3.48
SJT Completion Times (Time 1)		
Rate	12.28 ^A	3.47
Rank	14.75 ^B	3.63
Most/Least	13.29 ^C	3.22
SJT Completion Times (Time 2)		
Rate	8.84 ^A	2.45
Rank	10.42 ^B	2.98
Most/Least	8.98 ^A	3.08
FFM Completion Times		
Rate	3.73 ^A	1.12
Rank	22.23 ^B	9.86
Most/Least	6.67 ^C	1.96
Perceived Difficulty of SJT		
Rate	2.08 ^A	0.90
Rank	2.33 ^B	0.81
Most/Least	2.10 ^A	0.87
Preference for SJT Response Format		
Rate	3.33 ^A	1.48
Rank	2.56 ^B	1.38
Most/Least	3.77 ^C	1.45

Note. GMA scores range from 0-100; response distortion from 0-20; and SJT perceived difficulty and preference from 1-5. Neither the difficulty nor preference ratings were related to the Time 2 SJT (response format) condition. Completion times are reported in minutes. Where there are multiple rows for a variable, means with different superscripts are significantly different from each other ($p < .05$, one-tailed).

Evidence of the differential effect of response format on completion time was demonstrated with the personality measure as well. As shown in Table 2, just as with the SJT response formats, the rate-FFM took participants less time to complete ($M = 3.72$ minutes, $SD = 1.12$) than the most/least-FFM ($M = 6.67$, $SD = 1.96$, $t = 16.65$, $d = 1.85$, $p < .05$), which in turn took participants less time to complete than the rank-FFM ($M = 22.23$, $SD = 9.86$, $t = 19.70$, $d = 2.19$, $p < .05$).

Hypothesis 4 investigated the White–Hispanic subgroup differences of the SJT response formats. Table 3 indicates that mixed results were obtained for the differences in SJT scores between Whites and Hispanics. Specifically, Whites had higher mean scores than Hispanics for the Time 1 rate-SJT ($d = 0.43$) and Time 2 rank-SJT ($d = 0.38$), while Hispanics had higher mean scores than Whites for the Time 1 most/least-SJT ($d = -0.17$) and the Time 2 rate- ($d = -0.04$) and most/least-SJTs ($d = -0.22$). Interestingly, Whites and Hispanics obtained the exact same mean scores for the Time 1 rank-SJT ($d = 0.00$). Among these effects, only the difference for the Time 1 rate-SJT was statistically significant. Contrary to expectations, the pattern of results also shows that the subgroup differences for the rate-SJT (average of Time 1 and Time 2 $d = 0.20$), the rank-SJT (average of Time 1 and Time 2 $d = 0.19$), and the most/least-SJT (average of Time 1 and Time 2 $d = -0.20$) were all similar in magnitude. Thus, Hypotheses 4a and 4b were not supported. It should also be noted that Whites outperformed Hispanics on the rate- and rank-SJT, while Hispanics outperformed Whites on the most/least-SJT.

Concerning the sex-based subgroup differences for the SJT response formats (Hypothesis 5), the results presented in Table 4 indicate that the differences in SJT

scores for all three response formats favored women. Of note, women significantly outperformed men on the rank-SJT response format at both Time 1 and Time 2. The pattern of results shows that the rank-SJT displayed much larger subgroup differences (Time 1 $d = -0.55$, Time 2 $d = -0.41$, average $d = -.48$) than the most/least-SJT (Time 1 $d = -0.05$, Time 2 $d = -0.14$, average $d = -.10$), but that the most/least-SJT subgroup differences were also smaller than that of the rate-SJT (Time 1 $d = -0.22$, Time 2 $d = -0.04$, average $d = -.13$), providing support for Hypothesis 5a but not Hypothesis 5b. This pattern of results is the same as that obtained by Glaze et al. (2011).

Consistent with the random assignment of participants into conditions, there were negligible differences in response distortion between SJT response format conditions. Specifically, the results indicate that the response distortion scores for participants who completed the most/least-SJT ($M = 6.93$, $SD = 3.51$) were the same as those for participants who completed the rate-SJT ($M = 6.83$, $SD = 3.41$; $t = 0.26$, $d = 0.03$, $p > .05$), and the rank-SJT ($M = 6.70$, $SD = 3.46$; $t = 0.34$, $d = 0.04$, $p > .05$).

Table 1 presents the correlations between the SJT response formats and the response distortion scores. In reference to Hypotheses 6a and 6b, the results indicate that response distortion displayed positive relationships with the SJT scores, such that individuals who were likely responding in a socially desirable manner also had higher SJT scores. As expected, the rate-SJT displayed larger correlations with response distortion (Time 1 $r = .30$, Time 2 $r = .27$, average $r = .29$) than did the rank-SJT (Time 1 $r = .17$, Time 2 $r = .28$, average $r = .23$, $z = 0.57$, $p > .05$) and the most/least-SJT (Time 1 $r = .16$, Time 2 $r = .13$, average $r = .15$, $z = 1.33$, $p > .05$). Although these

differences were not statistically significant, the pattern of results demonstrated that individuals were more likely to engage in response distortion on a rate response format than a rank or most/least response format, and thus, Hypothesis 6a was supported. Contrary to expectations, the rank-SJT displayed larger, although not significantly different, correlations with response distortion than did the most/least-SJT ($z_r = 0.75, p > .05$), so Hypothesis 6b was not supported. These findings are consistent with Glaze et al.'s (2011) pattern of results, suggesting that the rate-SJT was the most susceptible to response distortion, although its susceptibility was comparable to that of the rank-SJT, and that the most/least-SJT was the least susceptible.

Hypotheses 7a and 7b posited that test-takers would have the most favorable reactions toward the rate-SJT, followed by the most/least-, and then the rank-SJT. Perceived difficulty ratings and mean preference ratings for the SJT response formats are reported in Table 2. The pattern of results indicate that the rank-SJT was rated as more difficult to complete than the most/least-SJT ($t = 2.45, d = 0.27, p < .05$), which in turn was rated only slightly more difficult than the rate-SJT ($t = 0.31, d = 0.02, p > .05$). In terms of preferences, the most/least-SJT was preferred over the rate-SJT ($t = 4.56, d = 0.30, p < .05$), which in turn was preferred over the rank-SJT ($t = 8.17, d = 0.54, p < .05$). Therefore, in general, the most/least-SJT engendered the most favorable reactions, closely followed by the rate-SJT; the rank-SJT was perceived least favorably. Because the difference in difficulty ratings between the rate- and most/least-SJTs was marginal, and the preference ratings were higher for the most/least-SJT than the rate-SJT, Hypothesis 7a was not supported. However, perceived difficulty for the rank-SJT was

greater than the most/least-SJT, and the most/least-SJT was preferred over the rank-SJT ($t = 12.98, d = 0.86, p < .05$), so Hypothesis 7b was supported.

Table 5 provides the descriptive statistics for the three SJT, GMA, and FFM response formats. Consistent with Glaze et al. (2011), at Time 1 the highest SJT scores were observed for the most/least-SJT, followed by the rate-SJT ($t = 17.27, d = 1.90, p < .05$), and then the rank-SJT ($t = 3.72, d = 0.41, p < .05$). Furthermore, a similar pattern of results was found at Time 2, such that the most/least-SJT had higher scores than the rate-SJT ($t = 16.99, d = 2.02, p < .05$), which in turn had higher scores than the rank-SJT ($t = 1.89, d = 0.22, p > .05$). Score distributions for the three SJT response formats at Time 1 can be found in Figure 2, and those at Time 2 can be found in Figure 3. The obtained scores indicate that it is easier to identify the most and least effective responses to an SJT scenario than to either rate or rank the responses, with the rank format being the most difficult. In addition, the rank-SJT displayed better distributional properties than the rate- and most/least-SJTs, in that it better approximates a normal distribution of scores. Lastly, the most/least-SJT displayed a rather large mean and the smallest standard deviation among the three response formats, and the data were also very negatively skewed, suggesting the possibility of range restriction with a lower variability in scores for this response format. These distributional properties were almost identical to those obtained by Glaze et al.

Table 3

White–Hispanic Subgroup Differences for the Integrity-Based Situational Judgment Test for All Response Formats

	TIME 1						TIME 2					
	Rate		Rank		Most/Least		Rate		Rank		Most/Least	
	White	Hispanic	White	Hispanic	White	Hispanic	White	Hispanic	White	Hispanic	White	Hispanic
<i>N</i>	115	28	101	33	114	25	103	24	110	25	115	37
Mean	61.55	55.04	54.46	54.46	83.34	84.40	58.36	59.00	55.26	50.20	84.02	85.57
<i>SD</i>	14.38	18.38	11.85	9.93	6.18	6.13	16.87	17.66	13.62	12.20	7.63	5.45
<i>d</i>	—	0.43*	—	0.00	—	-0.17	—	-0.04	—	0.38	—	-0.22

Note. Hispanics are compared to Whites such that a positive *d* indicates that Whites scored higher than Hispanics. * $p < .05$ (one-tailed).

Table 4

Sex-Based Subgroup Differences for the Integrity-Based Situational Judgment Test for All Response Formats

	TIME 1						TIME 2					
	Rate		Rank		Most/Least		Rate		Rank		Most/Least	
	Male	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male	Female
<i>N</i>	51	114	56	107	65	99	55	93	57	105	61	121
Mean	56.84	60.42	49.36	55.63	82.83	83.19	56.42	57.12	49.86	55.30	83.21	84.27
<i>SD</i>	17.48	15.78	12.80	10.56	7.48	6.06	18.73	18.06	14.02	13.06	7.33	7.44
<i>d</i>	—	-0.22	—	-0.55*	—	-0.05	—	-0.04	—	-0.41*	—	-0.14

Note. Females are compared to males such that a positive *d* indicates that males scored higher than females. * $p < .05$ (one-tailed).

Table 5
Descriptive Statistics for the Three Response Formats

Variable	Response Format					
	Rate		Rank		Most/Least	
	<i>Mean</i>	<i>SD</i>	<i>Mean</i>	<i>SD</i>	<i>Mean</i>	<i>SD</i>
SJT (Time 1)	59.32 ^A	16.35	53.47 ^B	11.72	83.05 ^C	6.64
SJT (Time 2)	56.86 ^A	18.25	53.38 ^B	13.61	83.92 ^C	7.40
GMA	39.51 ^A	7.27	25.24 ^B	11.33	76.42 ^C	7.65
Agreeableness	4.11 ^A	0.53	3.62 ^B	0.46	4.14 ^A	0.57
Conscientiousness	3.69 ^A	0.63	3.52 ^B	0.43	3.77 ^A	0.66
Emotional Stability	3.21	0.70	3.11	0.55	3.18	0.74

Note. SJT scores ranged from 0-100; GMA scores from 0-100; and FFM scores from 1-5. Means across rows with different superscripts are significantly different from each other ($p < .05$, two-tailed).

Table 6 presents the SJT reliability estimates. The SJT scores displayed high levels of internal consistency, with the rate-SJT having the highest reliability estimate (Time 1 $\alpha = .94$, Time 2 $\alpha = .95$, average $\alpha = .95$), followed by the rank-SJT (Time 1 $\alpha = .76$, Time 2 $\alpha = .81$, average $\alpha = .79$), and then the most/least-SJT (Time 1 $\alpha = .69$, Time 2 $\alpha = .76$, average $\alpha = .73$). The test-retest reliabilities in Table 6 also indicate that the rate-SJT had higher levels of reliability (.69) than the most/least-SJT (.64), which was in turn higher than the rank-SJT (.59). Alternate-form reliabilities were highest for the rank- and most/least-SJT (.67 and .70), followed by the rate- and rank-SJT (.47 and .35), and then the rate- and most/least-SJT (.29 and .40). As previously noted, the retest interval length varied across participants (i.e., 5-9 days between Time 1 and Time 2 sessions).

However, subsequent analyses presented in Appendix E demonstrated that the obtained results were not influenced by retest interval length.

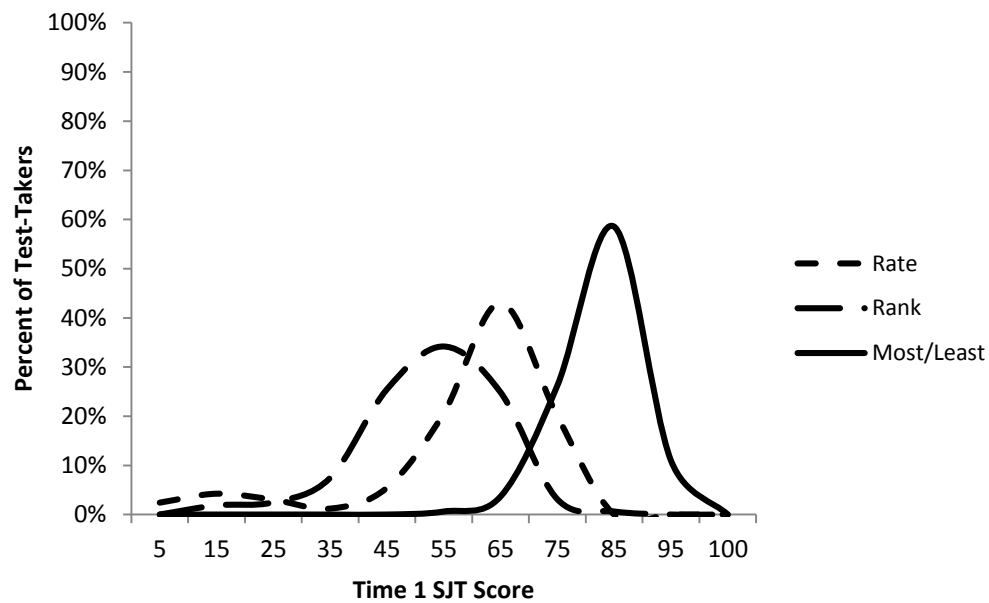


Figure 2. Time 1 score distributions for the three SJT response formats.

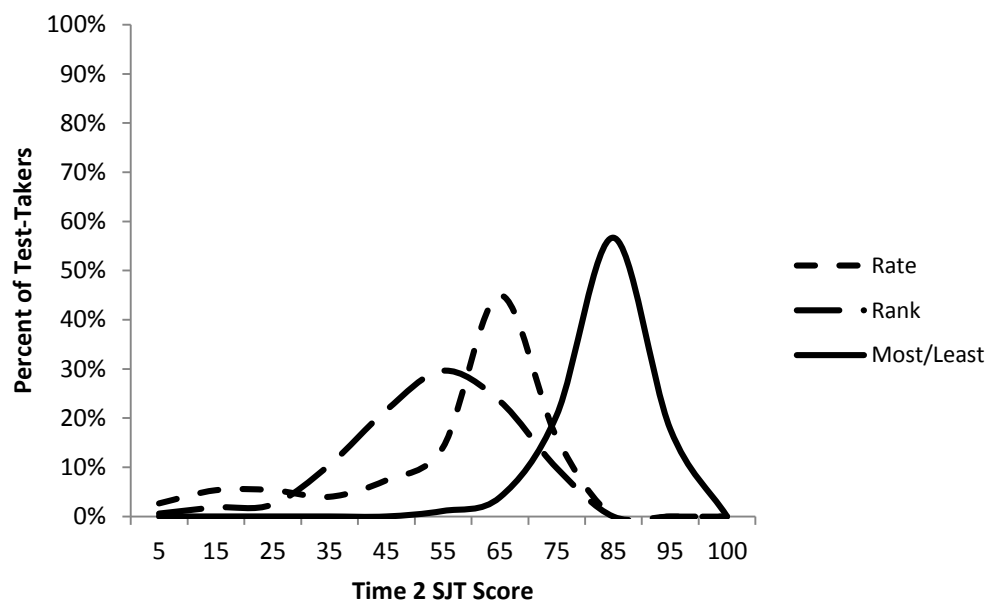


Figure 3. Time 2 score distributions for the three SJT response formats.

Table 6
Internal Consistency, Test-Retest, and Alternate-Form Reliabilities for the Integrity-Based Situational Judgment Test Scores

		TIME 1			Coefficient
		Test-Retest and Alternate-Form Reliabilities			
		Rate	Rank	Most/Least	
TIME 2	Rate	.69	.47 ^{A1}	.29 ^{B1}	.95
	Rank	.35 ^{A2}	.59	.67 ^{C1}	.81
	Most/Least	.40 ^{B2}	.70 ^{C1}	.64	.76
	Coefficient Alpha	.94	.76	.69	

Note. Retest interval = 5-9 days, $M = 6.99$, $SD = 0.74$, and 80.3% of the sample had a retest interval of 7 days. Test-retest reliabilities are on the diagonal. The length of the retest interval was not related to the retest condition (i.e., Time 2 response format condition; see Appendix E). Superscripts indicate alternate-form reliability pairs such that for instance ^{A1} denotes the rank/rate and ^{A2} the rate/rank alternate-form reliabilities.

DISCUSSION AND CONCLUSIONS

The objective of the present study was to undertake a constructive replication and extension of Glaze et al. (2011). A number of summary statements can be made on the basis of the obtained results. First, the results indicated that the relationship between SJT scores and GMA varied as a function of response format. Because the SJT used in the present study assessed a noncognitive construct (i.e., integrity), its scores should display a weak or zero correlation with GMA. However, consonant with Glaze et al.'s results, the posited greater information-processing demands and cognitive load associated with the most/least-, and to a greater extent the rank-SJT, resulted in stronger relationships with GMA than that of the rate-SJT, regardless of the GMA response format. Consequently, these results failed to provide support for the shared-common-response-method effect as an alternative explanation of the observed pattern of relationships because the matched pair response formats of the SJT and GMA test did not display the strongest correlations compared to the mismatched pair response formats.

Although the relationships between SJT scores and the specified FFM personality variables (i.e., agreeableness, conscientiousness, and emotional stability) did not unambiguously match Glaze et al.'s effects, the patterns of results were more in accord with the differences-in-g-loading explanation than they were with the shared-common-response-method explanation. Specifically, the rate-SJT generally displayed stronger correlations with FFM scores than the rank- and most/least-SJTs regardless of the FFM response format.

Also consistent with Glaze et al.'s results, the SJT response formats were associated with different completion times. Namely, the rank-SJT took participants longer to complete than the most/least-SJT, which in turn took longer than the rate-SJT, again a consequence of the posited differential information-processing and cognitive demands of the response formats. Interestingly, the rank response format of the personality measure also took participants longer to complete than the most/least response format, which in turn took participants longer to complete than the rate response format. Because this finding was consistent across selection tools, it suggests that in the context of noncognitive constructs, the cognitive loading effects engendered by the different response formats may be construct invariant. In summary, the present study provided additional evidence of the construct-related validity of the three SJT response formats, consonant with Glaze et al.'s findings.

The second summary statement is that the results pertaining to White–Hispanic subgroup differences for the three response formats were mixed. Specifically, Whites and Hispanics outperformed each other on different SJT response formats, although the only significant difference occurred for the rate-SJT, in which scores were higher for Whites. Furthermore, the magnitudes of the subgroup differences, regardless of their direction, were similar for all three formats. These findings appear to contradict that of past research. The extant literature has demonstrated that cognitively loaded tests display race-based subgroup differences favoring Whites (Roth et al., 2001) and that SJTs with higher cognitive loading increase the subgroup differences between Whites and Hispanics (Whetzel et al., 2008). Consistent with this, Glaze et al. (2011) reported

differences in SJT scores between Whites and Hispanics, and that these subgroup differences were largest for the rank-SJT and smallest for the rate-SJT. It is important to note that the White–Hispanic comparisons in the present study were unbalanced, as they were based on relatively small Hispanic *ns*, so this may limit the interpretation of these findings, particularly because they stand in contrast to past research.

Consonant with Glaze et al. (2011), the present study found that women scored higher on the SJT than men for all three response formats, particularly on the rank-SJT. Hence, the sex-based subgroup differences were much larger for the rank-SJT than the rate- and most/least-SJTs, which displayed similar, nonsignificant differences.

Third, concerning the relationship between SJT scores and response distortion, consistent with Glaze et al. (2011), the pattern of results indicated that the rate-SJT displayed the largest correlations with response distortion scores, followed by the rank-, and then the most/least-SJT. Because participants were randomly assigned to the SJT response format study conditions, it was not only expected that each group would display similar response distortion scores, but also that any differences in the relationship between SJT scores and response distortion would occur as a function of the response formats. Specifically, the rate-SJT allows participants to respond in a socially desirable manner because it does not impose discrimination and forced-choices between the response alternatives. Meanwhile, the most/least-SJT, and to a greater extent the rank-SJT, do not permit ties and force participants to weigh the relative effectiveness of each response option within an item. This highlights the influence that response formats can have on item responding and the psychometric properties of the scores obtained.

The fourth summary statement is that participants perceived the rank-SJT to be the most difficult to complete, followed by the most/least-, and then the rate-SJT. In addition, the most/least-SJT was preferred by participants over the rate-SJT, which in turn was preferred over the rank-SJT. Therefore, overall, the most/least response format received the most favorable and the rank response format received the least favorable test-taker reactions.

Lastly, consistent with Glaze et al.'s (2011) results, the highest SJT mean scores were observed for the most/least response format, followed by the rate, and then the rank response formats. However, as previously noted, the score distributions of the response formats suggest that ascertaining the correct answers for the most/least-SJT may be quite easy, and that this response format may be susceptible to range restriction. Conversely, the rank-SJT displayed better distributional properties and better approximated a normal distribution of scores.

Concerning the reliabilities of the response formats, the rate-SJT displayed higher internal consistency and test-retest reliability estimates than the rank- and most/least-SJTs, indicating that it displayed the highest levels of inter-item and temporal stability. To date, few studies in the extant literature have reported SJT test-retest reliabilities (e.g., Becker, 2004; Ployhart, Porr, & Ryan, 2004; Schmitt & Chan, 2006), and the present study was the first to report test-retest reliability estimates for the three response formats. Ployhart and Ehrhart (2003) were the first to report SJT test-retest reliabilities, which they found to range from .20 to .92 among a variety of instruction methods. More recently, Catano, Brochu, and Lamerson (2012) reported reliabilities of

.82 and .69 from two study samples. However, neither of these studies reported the constructs assessed, which limits the interpretation of their findings. Consequently, it is possible that the differences in the obtained test-retest reliabilities between the present study and past research may be due to the different constructs assessed by the respective SJTs, which is further complicated by the inability to make these comparisons when the constructs assessed are unknown. Another potential source of these differences is the retest interval length. The present study retested participants 5-9 days after the first SJT administration, while past research designs utilized longer retest intervals, such as two weeks (Catano et al., 2013) and “several weeks” (Ployhart & Ehrhart, 2003, p. 6). Thus, the differences in retest interval length may influence the test-retest reliability estimates. Although the present study demonstrated that interval length did not influence the patterns of results, this was based on relatively small differences in interval length. Therefore, an investigation of the magnitude of the effect that may result from larger differences in retest interval length is warranted. Finally, the highest alternate-form reliabilities were observed for the rank- and most/least-SJTs, which is likely due to the increased information-processing and cognitive demands associated with these response formats.

Scientific and Practical Implications

The scientific implications of the present study pertain to the unintended effect that different cognitive and information processing demands engendered by specified test design characteristics might have on the construct-related validity of tests by introducing construct-irrelevant variance. Hence, researchers should not be exclusively

method-focused, but rather pay closer attention to design features and the constructs measured by their tests.

The practical implications of the present study are germane to providing guidance and recommendations concerning the use of three common response formats and the boundary conditions under which a specified format may be preferable to others when administering SJTs that assess noncognitive constructs. Based on the results of their work, Glaze et al. (2011) provided initial recommendations and guidance for the use of different SJT response formats in organizational practice; specifically, they suggested that for noncognitive SJTs, the rate response format may be the preferred option. The results of the present study provide additional support for these conclusions and echo Glaze et al.'s recommendations. Furthermore, the present findings demonstrate that, in the context of noncognitive constructs, the use of the rate response format translates into higher levels of construct-related validity than the use of the rank or most/least response formats.

Limitations and Future Directions

There are limitations associated with the present study. First, conditions were simulated in which a personality measure had a response format similar to a GMA test (i.e., multiple-choice) and a GMA test with a response format similar to a personality measure (i.e., Likert response scale), which may appear unnatural to test-takers with no prior exposure to these formats. However, although the distributional assessment response format that was used with the personality measure is not common in

personality assessment, it is not conceptually unusual (cf. Edwards & Woehr, 2007), and can be used to abate response distortion.

It is also important to acknowledge that using the rate response format with a GMA test (i.e., APM) is potentially problematic. Because these items are objective, containing a singular correct answer, some individuals may reasonably hold the view that all response options other than the correct answer are, by definition, equally wrong. Consequently, they may rate the correct answer highly and all other response options low and with little variance. This event can impact the mean ratings of these items and influence the psychometric properties of the test scores. However, acknowledging that this method was suboptimal, it was the only way the pertinent issues of the present study could have been investigated. Specifically, the response formats were developed in order to demonstrate the patterns of relationships that the SJT scores share with GMA and the specified personality variables, and as such, permitted a comparative test of the two competing explanations for Glaze et al.'s (2011) observed effects.

Future research could also investigate the issues addressed here using a cognitively-based SJT. Specifically, one would expect an SJT assessing a cognitive construct to display strong relationships with GMA regardless of the response format. However, it is less clear how the information-processing and cognitive demands of the rank and most/least response formats will influence the differential magnitudes of these relationships, compared to that of the rate response format.

Future research might also examine the issues addressed here using other SJT response formats. For example, the “acceptable versus not acceptable” format requires

respondents to independently evaluate each response option within an item in terms of whether or not it reflects a satisfactory course of action in response to the scenario presented in the item stem. Thus, respondents assign a “yes/no” answer to each response option, such that they endorse the response options deemed to be acceptable answers and do not endorse those deemed to be unacceptable answers. To date, no studies in the literature have tested the efficacy of this format.

Finally, the posited differences-in-g-loading explanation is based on the premise that the rank and most/least response formats engender greater cognitive load and information processing demands than the rate response format. In accordance, these effects were investigated by looking at the relationships between GMA and the response formats, which is conceptually sound, given the theoretical and empirical relationship between GMA and working memory. That being said, GMA was nevertheless used as a proxy for working memory. Hence, future research could directly investigate these effects by administering a working memory test. It is likely that the patterns of relationships among the response formats and working memory will be similar to that of GMA due to the strong relationship between GMA and working memory ($\rho = .479$; Ackerman, Beier, & Boyle, 2005). Furthermore, the magnitudes of the response format–working memory correlations might be stronger than the response format–GMA correlations, given the posited information processing requirements of the rank and most/least response formats. Specifically, working memory should influence test-takers’ ability to accurately and efficiently make the relative judgments among the response options for these formats.

Conclusions

The present study undertook a constructive replication of Glaze et al. (2011) by comparing three SJT response formats (i.e., rate, rank, and most/least) in terms of construct-related validity, subgroup differences, and susceptibility to response distortion. This work also extended past research assessing test-taker reactions to, and the internal psychometric properties of the three response formats. The results obtained here contribute to the literature by demonstrating that in the context of noncognitive constructs, the response format used when administering an SJT matters. In so doing, the present study provided additional support for the differences-in-g-loading explanation of the relationships that the SJT response format scores share with GMA and the specified personality variables (i.e., agreeableness, conscientiousness, and emotional stability). Furthermore, based on these findings, it appears that the rate response format is superior to the rank and most/least response formats, and in most cases should be the preferred response format for selection test developers.

REFERENCES

- Ackerman, P. L., Beier, M. E., & Boyle, M. O. (2005). Working memory and intelligence: The same or different constructs? *Psychological Bulletin*, 131, 30-60. doi: 10.1037/0033-2909.131.1.30.
- Anderson, N. (2004). The dark side of the moon: Applicant perspectives, negative psychological effects (NPEs), and candidate decision making in selection. *International Journal of Selection and Assessment*, 12, 1-8.
- Arthur, W. Jr., & Day, D. V. (1994). Development of a short form for the Raven Advanced Progressive Matrices Test. *Educational and Psychological Measurement*, 54, 394-403.
- Arthur, W. Jr., & Villado, A. J. (2008). The importance of distinguishing between constructs and methods when comparing predictors in personnel selection research and practice. *Journal of Applied Psychology*, 93, 435-442.
- Arthur, W. Jr., Doverspike, D., Barrett, G. V., & Miguel, R. (2013). Chasing the Title VII holy grail: The pitfalls of guaranteeing adverse impact elimination. *Journal of Business and Psychology*. Online first publication. doi: 10.1007/s10869-013-9289-6
- Arthur, W. Jr., Tubre, T. C., Paul, D. S., & Sanchez-Ku, M. L. (1999). College-sample psychometric and normative data on a short form of the Raven Advanced Progressive Matrices Test. *Journal of Psychoeducational Assessment*, 17, 354-361.

- Basilli, J. N., & Scott, B. S. (1996). Response latency as a signal to question problems in survey research. *Public Opinion Quarterly*, 60, 390-399.
- Bauer, T. N., & Truxillo, D. M. (2006). Applicant reactions to situational judgment tests: Research and related practical issues. In J. A. Weekley, & R. E. Ployhart (Eds.), *Situational judgment tests: Theory, measurement, and application* (pp. 223-249). Mahwah, NJ: Lawrence Erlbaum Associates.
- Becker, T. E. (2004). *Development and validation of a scenario-based measure of employee integrity*. Paper presented at the 19th Annual Convention of the Society for Industrial and Organizational Psychology. Chicago, IL.
- Bergman, M. E., Drasgow, F., Donovan, M. A., Henning, J. B., & Juraska, S. E. (2006). Scoring situational judgment tests: Once you get the data, your troubles begin. *International Journal of Selection and Assessment*, 14, 223-235.
- Berry, C. M., Sackett, P. R., & Wiemann, S. (2007). A review of recent developments in integrity test research. *Personnel Psychology*, 60, 271-301.
- Berry, D. T., Wetter, M. W., Baer, R. A., Larsen, L., Clark, C., & Monroe, K. (1992). MMPI-2 random responding indices: Validation using a self-report methodology. *Psychological Assessment*, 4, 340-345.
- Bond, J. A. (1986). Inconsistent responding to repeated MMPI items: Is its major cause really carelessness? *Journal of Personality Assessment*, 50, 50-64.
- Bond, J. A. (1987). The process of responding to personality items: Inconsistent responses to repeated presentation of identical items. *Personality and Individual Differences*, 8, 409-417.

- Bradshaw, J. (1990). Test-takers' reactions to a placement test. *Language Testing*, 7, 13-30.
- Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices Test. *Psychological Review*, 97, 404-431.
- Catano, V. M., Brochu, A., & Lamerson, C. D. (2012). Assessing the reliability of situational judgment tests used in high-stakes situations. *International Journal of Selection and Assessment*, 20, 334-346.
- Chan, D., & Schmitt, N. (1997). Video-based versus paper-and-pencil method of assessment in situational judgment tests: Subgroup differences in test performance and face validity perceptions. *Journal of Applied Psychology*, 82, 143-159.
- Chan, D., & Schmitt, N. (2002). Situational judgment and job performance. *Human Performance*, 15, 233-254.
- Chan, D., & Schmitt, N. (2004). An agenda for future research on applicant reactions to selection procedures: A construct-oriented approach. *International Journal of Selection and Assessment*, 12, 9-23.
- Christian, M. S., Edwards, B. D., & Bradley, J. C. (2010). Situational judgment tests: Constructs assessed and a meta-analysis of their criterion-related validities. *Personnel Psychology*, 63, 83-117.

- Clevenger, J., Pereira, G. M., Wiechmann, D., Schmitt, N., & Harvey, V. S. (2001). Incremental validity of situational judgment tests. *Journal of Applied Psychology, 86*, 410-417.
- Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*, 98-104.
- Cullen, M. J., Sackett, P. R., & Lievens, F. (2006). Threats to the operational use of situational judgment tests in the college admission process. *International Journal of Selection and Assessment, 14*, 142-155.
- Edwards, A. L. (1957). *The social desirability variable in personality assessment and research*. Ft. Worth, TX: Dryden Press.
- Edwards, B. D., & Woehr, D. J. (2007). An examination and evaluation of frequency-based personality measurement. *Personality and Individual Differences, 43*, 803-814.
- Erdfelder, E., Faul, F., & Buchner, A. (1996). GPOWER: A general power analysis program. *Behavior Research Methods, Instruments, & Computers, 28*, 1-11.
- Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analysis using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, Instruments, & Computers, 41*, 1149-1160.
- Glaze, R. M., Jarrett, S., Schurig, I., Arthur, W. Jr., & Taylor, J. E. (2011). *The efficacy of three situational judgment test (SJT) response formats*. Paper presented at the 26th Annual Conference of the Society for Industrial and Organizational Psychology. Chicago, IL.

- Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment, 4*, 26-42.
- Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. In I. Mervielde, I. Deary, F. De Fruyt, & F. Ostendorf (Eds.), *Personality psychology in Europe, Vol. 17* (pp. 7-28). Tilburg, The Netherlands: Tilburg University Press.
- Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. G. (2006). The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality, 40*, 84-96.
- Guilford, J. S., Zimmerman, W. S., & Guilford, J. P. (1976). *The Guilford-Zimmerman Temperament Survey Handbook: Twenty-five years of research and application*. San Diego, CA: EdITS.
- Hand, J. (1964). Measurement of response sets. *Psychological Reports, 14*, 907-913.
- Konstabel, K., Aavik, T., & Allik, J. (2006). Social desirability and consensual validity of personality traits. *European Journal of Personality, 20*, 549-566.
- Kurtz, J. E., & Parrish, C. L. (2001). Semantic response consistency and protocol validity in structured personality assessment: The case of the NEO-PI-R. *Journal of Personality Assessment, 76*, 315-322.
- McDaniel, M. A., & Nguyen, N. T. (2001). Situational judgment tests: A review of practice and constructs assessed. *International Journal of Selection & Assessment, 9*, 103-113.

- McDaniel, M. A., Morgeson, F. P., Finnegan, E. B., Campion, M. A., & Braverman, E. P. (2001). Use of situational judgment tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology*, 86, 730-740.
- McDaniel, M. A., Whetzel, D. L., Hartman, N. S., Nguyen, N. T., & Grubb, W. L. (2006). Situational judgment tests: Validity and an integrative model. In J. A. Weekley, & R. E. Ployhart (Eds.), *Situational judgment tests: Theory, measurement, and application* (pp. 183-203). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.
- Motowidlo, S. J., Dunnette, M. D., & Carter, G. W. (1990). An alternative selection procedure: The low-fidelity simulation. *Journal of Applied Psychology*, 75, 640-647.
- Nguyen, N. T., Biderman, M. D., & McDaniel, M. A. (2005). Effects of response instructions on faking a situational judgment test. *International Journal of Selection & Assessment*, 13, 250-260.
- Olsen-Buchanan, J. B., Drasgow, F., Moberg, P. J., Mead, A. D., Keenan, P. A., & Donovan, M. A. (1998). Interactive video assessment of conflict resolution skills. *Personnel Psychology*, 51, 1-24.
- Ones, D. S. (1993). The construct validity of integrity tests. (Doctoral Dissertation). Retrieved from ProQuest Information & Learning (1996-73265-001).
- Paulhus, D. L. (1991a). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (pp. 17-59). San Diego, CA: Academic Press.

- Paulhus, D. L. (1991b). Balanced Inventory of Desirable Responding (BIDR) reference manual for version 6. (*Manual available from author at Department of Psychology, University of British Columbia, Vancouver, B. C., Canada V6T1Y7.*).
- Paulhus, D. L. (1994). *Balanced Inventory of Desirable Responding: Reference manual for BIDR Version 6*. Unpublished manuscript, University of British Columbia, Vancouver, Canada.
- Paulhus, D. L. (2002). Socially desirable responding: The evolution of the construct. In H. Braun, D. Jackson, & D. Wiley (Eds.), *The role of constructs in psychological and educational measurement* (pp. 49-69). Mahwah, NJ: Lawrence Erlbaum Associates.
- Ployhart, R. E., & Ehrhart, M. G. (2003). Be careful what you ask for: Effects of response instructions on the construct validity and reliability of situational judgment tests. *International Journal of Selection & Assessment, 11*, 1-16.
- Ployhart, R. E., & Holtz, B. C. (2008). The diversity-validity dilemma: Strategies for reducing racioethnic and sex subgroup differences and adverse impact in selection. *Personnel Psychology, 61*, 153-172.
- Ployhart, R. E., & MacKenzie, W. I. (2011). Situational judgment tests: A critical review and agenda for the future. In S. Zedeck (Ed.), *APA handbook of industrial and organizational psychology, Vol. 2: Selecting and developing members for the organization* (pp. 237-252). Washington, DC: APA.

- Ployhart, R. E., Porr, W., & Ryan, A. M. (2004). *New developments in SJTs: Scoring, coaching, and incremental validity*. Paper presented as the 19th Annual Convention of the Society for Industrial and Organizational Psychology. Chicago, IL.
- Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology, 88*, 879-903.
- Raven, J. (1989). The Raven Progressive Matrices: A review of national norming studies and ethnic and socioeconomic variation within the United States. *Journal of Educational Measurement, 26*, 1-16.
- Raven, J. (2000). The Raven's Progressive Matrices: Change and stability over culture and time. *Cognitive Psychology, 41*, 1-48.
- Rosse, J. G., Stecher, M. D., Miller, J. L., & Levin, R. A. (1998). The impact of response distortion on preemployment personality testing and hiring decisions. *Journal of Applied Psychology, 83*, 634-644.
- Roth, P. L., Bevier, C. A., Bobko, P., Switzer, F. S., & Tyler, P. (2001). Ethnic group differences in cognitive ability in employment and educational settings: A meta-analysis. *Personnel Psychology, 54*, 297-330.
- Ryan, M. A., & Ployhart, R. E. (2000). Applicants' perceptions of selection procedures and decisions: A critical review and agenda for the future. *Journal of Management, 26*, 565-606.

- Schmitt, N., & Chan, D. (2006). Situational judgment tests: Method or construct? In J. A. Weekley, & R. E. Ployhart (Eds.), *Situational judgment test: Theory, measurement, and application* (pp. 135-155). Mahwah, NJ: Lawrence Erlbaum Associates.
- Shyamsunder, A., & McCune, E. A. (2009). *Test-taker reactions to item formats used in online selection assessments*. Paper presented at the 24th Annual Conference of the Society for Industrial and Organizational Psychology. Atlanta, GA.
- Stober, J., & Dette, D. E. (2002). Comparing continuous and dichotomous scoring of the Balanced Inventory of Desirable Responding. *Journal of Personality Assessment*, 78, 370-389.
- Truxillo, D. M., Seitz, R., & Bauer, T. N. (2008). The role of cognitive ability in self-efficacy and self-assessed test performance. *Journal of Applied and Social Psychology*, 38, 903-918.
- Weekley, J. A., & Jones, C. (1997). Video-based situational testing. *Personnel Psychology*, 50, 25-49.
- Wetter, M. W., Baer, R. A., Berry, D. T., Smith, G. T., & Larsen, L. H. (1992). Sensitivity of MMPI-2 validity scales to random responding and malingering. *Psychological Assessment*, 4, 369-374.
- Whetzel, D. L., McDaniel, M. A., & Nguyen, N. T. (2008). Subgroup differences in situational judgment test performance: A meta-analysis. *Human Performance*, 21, 291-309.

- Yan, T., & Tourangeau, R. (2008). Fast times and easy questions: The effect of age, experience, and question complexity on web survey response times. *Applied Cognitive Psychology*, 22, 51-68.
- Zerbe, W. J., & Paulhus, D. L. (1987). Socially desirable responding in organizational behavior: A reconception . *Academy of Management Review*, 12, 250-264.

APPENDIX A

TEST BATTERY SCALE SAMPLE ITEMS

SITUATIONAL JUDGMENT TEST (SJT) - INTEGRITY SAMPLE ITEM (RATE RESPONSE FORMAT)

This measure describes 20 hypothetical workplace scenarios. For each scenario, please rate **each** action in terms of its effectiveness as a solution or response to the situation presented in the scenario. Please use the rating scale below:

Highly Ineffective Action or Response	Ineffective Action or Response	Neither Effective nor Ineffective Action or Response	Effective Action or Response	Highly Effective Action or Response
1	2	3	4	5

Place an “X” in the appropriate box of the effectiveness scale for **each** action or response to the scenario.

An employee is hired into a situation where it is common for all employees to take things from work. Rate the effectiveness of each of the following in terms of what you think a new hire should do.		1	2	3	4	5
1.	Take something small to be sure they fit into the work culture.					
2.	Take products from work only when they really need extra cash.					
3.	Write an anonymous letter notifying the corporate office of the situation.					
4.	Take things and give them to friends and family members to help them out.					
5.	Work through other employees to sell products and split the cash.					

SJT SAMPLE ITEM (RANK RESPONSE FORMAT)

This measure describes 20 hypothetical workplace scenarios. For each scenario, please rank each action in terms of its effectiveness as a solution or response to the situation presented in the scenario. Assign your ranks such that **1 = MOST effective** and **5 = LEAST effective**. There cannot be any ties in the ranks.

An employee is hired into a situation where it is common for all employees to take things from work. Rank the effectiveness of each of the following in terms of what you think a new hire should do.		
		Rank
1.	Take something small to be sure they fit into the work culture.	
2.	Take products from work only when they really need extra cash.	
3.	Write an anonymous letter notifying the corporate office of the situation.	
4.	Take things and give them to friends and family members to help them out.	
5.	Work through other employees to sell products and split the cash.	

SJT SAMPLE ITEM (MOST/LEAST RESPONSE FORMAT)

This measure describes 20 hypothetical workplace scenarios. For each scenario, please select the **MOST effective** and **LEAST effective** actions in terms of their effectiveness as solutions or responses to the situation presented in the scenario. That is, for each set of 5 actions that are presented for each scenario, select the one that you think is the **MOST effective** and the one that you think is the **LEAST effective**.

Place an **"M"** in the space next to the **most effective** action, and place an **"L"** in the space next to the **least effective** action. Leave the spaces next to the other three actions blank.

An employee is hired into a situation where it is common for all employees to take things from work. Identify the most effective and least effective of the following in terms of what you think a new hire should do.		
1.	Take something small to be sure they fit into the work culture.	
2.	Take products from work only when they really need extra cash.	
3.	Write an anonymous letter notifying the corporate office of the situation.	
4.	Take things and give them to friends and family members to help them out.	
5.	Work through other employees to sell products and split the cash.	

**GENERAL MENTAL ABILITY (GMA) -
 RAVEN’S ADVANCED PROGRESSIVE MATRICES
 GMA SAMPLE ITEM (RATE VERSION)²**

The top part of each item is a pattern with a piece cut out of it. Below the pattern are eight pieces that potentially complete the pattern. The piece with the best fit will complete the pattern correctly both horizontally and vertically. Look at the pattern and rate (using the 5-point scale that is provided) **HOW WELL** each of the eight

Item Pattern

**8 Response
 Alternatives**

Using the scale below, please rate how well each piece completes the pattern:

Very Poor Fit		Very Good Fit		
1	2	3	4	5

1	①	②	③	④	⑤
2	①	②	③	④	⑤
3	①	②	③	④	⑤
4	①	②	③	④	⑤
5	①	②	③	④	⑤
6	①	②	③	④	⑤
7	①	②	③	④	⑤
8	①	②	③	④	⑤

² The item pattern and response alternatives could not be presented here due to copyright protection of the APM

GMA SAMPLE ITEM (RANK VERSION)³

The top part of each item is a pattern with a piece cut out of it. Below the pattern are eight pieces that potentially complete the pattern. The piece with the best fit will complete the pattern correctly both horizontally and vertically. Look at the pattern and rank the eight pieces on **HOW WELL each completes the pattern**.

Item Pattern

**8 Response
Alternatives**

Using the spaces below, please rank all eight pieces in order of how well they complete the pattern, assigning '1' to the piece that best fits the pattern and '8' to the piece that is the worst fit.

1	
2	
3	
4	
5	
6	
7	
8	

³ The item pattern and response alternatives could not be presented here due to copyright protection of the APM

GMA SAMPLE ITEM (MOST/LEAST VERSION)⁴

The top part of each item is a pattern with a piece cut out of it. Below the pattern are eight pieces that potentially complete the pattern. The piece with the best fit will complete the pattern correctly both horizontally and vertically. Look at the pattern and identify the piece that **BEST** completes the pattern and the piece that is the **WORST** fit.

Item Pattern

**8 Response
Alternatives**

Please identify the piece that **BEST** completes the pattern and the piece that is the **WORST** fit. Using the spaces below, place a 'B' next to the **BEST** fit and a 'W' next to the **WORST** fit.

1	
2	
3	
4	
5	
6	
7	
8	

⁴ The item pattern and response alternatives could not be presented here due to copyright protection of the APM

PERSONALITY (FIVE FACTOR MODEL) - INTERNATIONAL PERSONALITY ITEM POOL (IPIP)

FFM SAMPLE ITEMS (RATE RESPONSE FORMAT)

Listed below are phrases describing people's behaviors. Please use the scale provided below to identify how accurately each statement describes *you*. Describe yourself as you generally are now, not as you wish to be in the future. Describe yourself as you honestly see yourself, in relation to other people you know of the same sex as you are, and roughly your same age. Please read each statement carefully, and then bubble in the number that corresponds to the number on the scale below.

Very Inaccurate	Inaccurate	Neither Inaccurate nor Accurate	Accurate	Very Accurate
1	2	3	4	5

1.	The life of the party	① ② ③ ④ ⑤
2.	Feel little concern for others	① ② ③ ④ ⑤

FFM SAMPLE ITEMS (RANK RESPONSE FORMAT)

Listed below are phrases describing people's behaviors. Please use the scale provided below to identify how accurately each statement describes **you**. Describe yourself as you generally are now, not as you wish to be in the future. Describe yourself as you honestly see yourself, in relation to other people you know of the same sex as you are, and roughly your same age. Please read each statement carefully, and then assign a percentage to **each** level in the scale that reflects how often it is true of you for the item statement. The percentages you assign to the five levels must equal 100% and **THERE CAN BE NO TIES**, so the level that is descriptive of you most often for each item will receive the largest percentage, and the level that is descriptive of you least often for each item will receive the smallest percentage. Please read the example item first, and then answer all items in the same manner using the scale below:

		Very Inaccurate	Inaccurate	Neither Inaccurate nor Accurate	Accurate	Very Accurate	
1.	The life of the party	_____	+	_____	+	_____	= 100
2.	Feel little concern for others	_____	+	_____	+	_____	= 100

FFM SAMPLE ITEMS (MOST/LEAST RESPONSE FORMAT)

Listed below are phrases describing people's behaviors. Please use the scale provided below to identify how accurately each statement describes **you**. Describe yourself as you generally are now, not as you wish to be in the future. Describe yourself as you honestly see yourself, in relation to other people you know of the same sex as you are, and roughly your same age. Please read each statement carefully, and then identify the level on the scale that **MOST accurately** describes your behavior and the level that **LEAST accurately** describes your behavior. For **each** item, use an "**M**" to indicate the level of the scale that **MOST accurately** describes your behavior, **AND** an "**L**" for the level that **LEAST accurately** describes your behavior.

		Very Inaccurate	Inaccurate	Neither Inaccurate nor Accurate	Accurate	Very Accurate
1.	The life of the party	_____	_____	_____	_____	_____
2.	Feel little concern for others	_____	_____	_____	_____	_____

RESPONSE DISTORTION – BALANCED INVENTORY OF DESIRABILITY RESPONDING (BIDR) VERSION 6 FORM 20A SAMPLE ITEMS

Using the scale below, please rate how true each statement is in describing *you*.

Not True	Somewhat True				Very True
① -----	② -----	③ -----	④ -----	⑤ -----	⑥ ----- ⑦

		Not True	Somewhat True				Very True
1.	I sometimes tell lies if I have to.	①	②	③	④	⑤	⑥ ⑦
2.	I never cover up my mistakes.	①	②	③	④	⑤	⑥ ⑦

TEST-TAKER REACTIONS SAMPLE ITEMS

Using the scale below, please rate the extent to which you agree with the following three statements regarding the measure you just completed.

		Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree
		1	2	3	4	5
1.	The response format was difficult to use.	①	②	③	④	⑤
2.	I wish there was some other response format that I could have used.	①	②	③	④	⑤
3.	The response format was difficult to understand.	①	②	③	④	⑤

APPENDIX B

**PILOT TESTING FOR THE THREE RAVEN’S ADVANCED PROGRESSIVE
MATRICES RESPONSE FORMATS**

Test-takers are allowed 15 minutes to complete the standard short-form version of the Raven’s Advanced Progressive Matrices (APM). Specifically, the ‘standard’ version refers only to the response format used in the original APM test. However, more decisions must be made in responding to items on the three response formats developed for the present study. Thus, two rounds of pilot testing were conducted in order to determine the appropriate time limit for the three response formats. First, a community sample of ten individuals was recruited. Each pilot tester completed one of the three response formats and self-reported their completion time. Based on these self-reports, a time limit of 25 minutes was initially assigned, which was then used with nine students who self-selected into the present study via the Texas A&M University Subject Pool and completed one of the three response formats based on random assignment of study conditions. Three of the nine wave 2 pilot testers did not complete the test before the 25 minute time limit, so upon further consideration, the time limit was raised to 30 minutes. The justification for doing so was that the APM was designed to be a test of power, not a timed test, so it was psychometrically more beneficial to provide enough time for most participants to complete the test. The 30 minute time limit was then used for the actual data collection. The alternate-form reliabilities for the APM response formats at Time 1

with the standard short-form APM format at Time 2 were .52, .46, and .62 for the rate-, rank-, and most/least-GMA, respectively.

APPENDIX C
SCORING KEY FOR THE NORMATIVE DATA SCORING METHOD
OF THE GMA TEST

Alternative methods to the SME judgment approach to generating scoring keys were considered for the APM response formats. Thus, the SME consensus approach to generating the scoring keys was compared to APM normative data. In doing so, 908 cases of the standard short form APM response format were gathered from prior studies conducted within a Texas A&M University I/O psychology research lab, and frequencies of the response options selected within each item were obtained. These relative frequencies informed the rank ordering of response options for each item in the rank-GMA format, and the highest and lowest ranked response options were designated as the best and worst response options for the most/least-GMA format. Due to the characteristics of the rate response format, normative data could not be used as a comparison for rating the effectiveness of each response option.

Three instances occurred in which two response options within an item were selected equally frequently. To break these ties, the response options ranked higher by the SME panel for said items were assigned the higher ranking. Table C.1 presents a comparison of the descriptive statistics for the SME consensus and normative data scoring keys of the GMA test. The results indicate negligible differences in the means for the rank and most/least response formats between the consensus and normative data scoring keys. Furthermore, the scoring keys displayed high correlations for both the

rank (.87) and most/least (.95) response formats. This demonstrates that using a different approach to generating the scoring keys for the APM response formats would not have made a significant difference than using the consensus method, and thus, would have obtained the same pattern of results for the hypothesized effects.

Table C.1
Descriptive Statistics for the SME and Normative Data Scoring Keys of the GMA Test

Response Format	SME Consensus Ratings		Normative Data		<i>r</i>
	<i>Mean</i>	<i>SD</i>	<i>Mean</i>	<i>SD</i>	
Rate	39.51	7.27	—	—	
Rank	25.24	11.33	23.14	9.09	.87*
Most/Least	76.42	7.65	77.41	7.61	.95*

Note. * $p < .05$ (one-tailed). SJT scores ranged from 0-100.

APPENDIX D

ALTERNATE SCORING METHOD FOR THE RESPONSE FORMATS

An alternative scoring algorithm for the SJT and GMA test was implemented to further compare the scoring keys used in the present study analyses. Specifically, a partial credit scoring method developed by Glaze et al. (2011) was used for the rate and rank response formats, and a partial scoring method that is common in the literature (Motowidlo et al., 1990) was used for the most/least response format.

For the rate and rank response format partial scoring method, the absolute difference between the SME ratings and the test-taker responses were calculated for each response option, and this absolute deviation was summed across all response options for all items. The total score was then scaled to 100 for ease of interpretation. For the most/least format, test-takers were awarded a point for identifying the most and least effective response option (as identified by the SMEs). However, test-takers were not docked a point for identifying the most effective response option as the least effective response option, or identifying the least effective response option as the most effective.

As expected, different results were obtained by employing the alternate (i.e., partial) scoring method for the three response formats. Specifically, mean scores for the SJT and GMA test were higher, and correlations among the SJT and GMA test response formats changed somewhat. However, the scores obtained via the alternate scoring method displayed very high correlations with the standard scoring method, and thus, did not influence the overall patterns of results germane to the hypotheses of the present

study. Tables D.1 through D.6 present the results pertinent to the study hypotheses, using the alternate scoring methods.

Table D.1

Integrity-Based Situational Judgment Test Correlations with General Mental Ability for the Alternate Scoring Method of the Three Response Formats

	SJT Response Format					
	Rate		Rank		Most/Least	
	Time 1	Time 2	Time 1	Time 2	Time 1	Time 2
GMA Response Format						
Rate	<u>.04</u>	<u>-.20</u>	.32*	.44*	.03	.15
Rank	-.02	.04	<u>.36*</u>	<u>.28*</u>	.19	-.04
Most/Least	.11	.18	-.10	.20	<u>-.07</u>	<u>.15</u>

Note. Underlined correlations represent those that would be expected to be highest, relative to the others, if the results are indicative of a shared-common-response-method effect. * $p < .05$ (one tailed).

Table D.2

Descriptive Statistics for the Absolute and Partial Credit Scoring Methods of the Integrity-Based Situational Judgment Test and GMA Test

	Scoring Method				<i>r</i>
	Absolute		Partial		
	<i>Mean</i>	<i>SD</i>	<i>Mean</i>	<i>SD</i>	
SJT (Time 1)					
Rate	59.32 ^A	16.35	86.14 ^A	7.88	.89*
Rank	53.47 ^B	11.72	78.28 ^B	8.88	.95*
Most/Least	83.05 ^C	6.64	68.08 ^C	11.23	.99*
SJT (Time 2)					
Rate	56.86 ^A	18.25	84.67 ^A	9.47	.90*
Rank	53.38 ^B	13.61	77.67 ^B	10.79	.94*
Most/Least	83.92 ^C	7.40	69.52 ^C	12.75	.99*
GMA					
Rate	39.51 ^A	7.27	72.53 ^A	7.03	.66*
Rank	25.24 ^B	11.33	67.08 ^B	10.73	.89*
Most/Least	76.42 ^C	7.65	54.42 ^C	13.56	.99*

Note. * $p < .05$ (one-tailed). SJT scores ranged from 0-100 and FFM scores from 1-5. Within-variable means with different superscripts are significantly different from each other ($p < .05$, one tailed).

Table D.3

White–Hispanic Subgroup Differences Using the Alternate Scoring Method for the Integrity-Based Situational Judgment Test for All Response Formats

	TIME 1						TIME 2					
	Rate		Rank		Most/Least		Rate		Rank		Most/Least	
	White	Hispanic	White	Hispanic	White	Hispanic	White	Hispanic	White	Hispanic	White	Hispanic
<i>N</i>	115	28	101	33	114	25	103	24	110	25	115	37
Mean	87.64	84.43	78.94	79.39	68.47	70.00	85.57	85.80	78.91	75.08	69.87	71.82
<i>SD</i>	5.35	7.58	8.69	6.40	10.68	10.99	8.83	7.16	10.49	10.92	13.15	10.23
<i>d</i>	—	0.55*	—	-0.05	—	-0.14	—	-0.03	—	0.36*	—	-0.16

Note. Hispanics are compared to Whites such that a positive *d* indicates that Whites scored higher than Hispanics. **p* < .05 (one-tailed).

Table D.4

Sex-Based Subgroup Differences Using the Alternate Scoring Method for the Integrity-Based Situational Judgment Test for All Response Formats

	TIME 1						TIME 2					
	Rate		Rank		Most/Least		Rate		Rank		Most/Least	
	Male	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male	Female
<i>N</i>	51	114	56	107	65	99	55	92	57	105	60	121
Mean	85.10	86.60	75.54	79.72	67.96	68.16	84.35	84.86	75.04	79.09	68.25	70.15
<i>SD</i>	8.29	7.68	9.98	7.93	12.29	10.55	10.35	8.96	12.39	9.59	12.67	12.79
<i>d</i>	—	-0.19	—	-0.50*	—	-0.02	—	-0.05	—	-0.38*	—	-0.15

Note. Females are compared to males such that a positive *d* indicates that males scored higher than females. **p* < .05 (one-tailed).

Table D.5

Internal Consistency, Test-Retest, and Alternate-Form Reliabilities for the Alternate Scoring Method of the Integrity-Based Situational Judgment Test Scores

TIME 1					
Test-Retest and Alternate-Form Reliabilities					Coefficient Alpha
		Rate	Rank	Most/Least	
TIME 2	Rate	.51	.31 ^{A1}	.49 ^{B1}	.95
	Rank	.49 ^{A2}	.59	.71 ^{C1}	.87
	Most/Least	.35 ^{B2}	.68 ^{C1}	.62	.68
	Coefficient Alpha	.97	.92	.75	

Note. Retest interval = 5-9 days, $M = 6.99$, $SD = 0.74$, and 80.3% of the sample had a retest interval of 7 days. Test-retest reliabilities are in the diagonal. The length of the retest interval is not related to the retest condition (i.e., Time 2 response format condition; see Appendix E). Superscripts indicate alternate-form reliability pairs such that for instance ^{A1} denotes the rank/rate and ^{A2} the rate/rank alternate-form reliabilities.

Table D.6

Alternate-Form Reliabilities for the Alternate Scoring Method of the GMA Test

	Rate	Rank	Most/Least
Standard GMA	.52	.43	.63

APPENDIX E

TIME INTERVAL ANALYSIS

The time interval between sessions was not held constant. Specifically, participants were allowed to complete Time 2 between 5 and 9 days after completing Time 1 ($M = 6.99$, $SD = 0.74$; see Table E.1). Thus, additional analyses were conducted in order to ensure that interval length did not confound the results obtained. A single factorial analysis of variance (ANOVA) of the three SJT response formats on retest interval revealed no systematic differences between retest interval lengths ($F = 0.01$, $R^2 < .001$, $p > .05$), and all correlations between retest interval and the scores for all measures at Time 2 were small (i.e., less than .10) and not significant. These results indicate that the variation in time interval lengths between study sessions did not differentially impact participant scores at Time 2.

Table E.1
Retest Interval Length Frequencies

Interval Length	Frequency	Percent
5	31	6.3
6	17	3.5
7	395	80.3
8	22	4.5
9	27	5.5

Note. Retest interval lengths reported in days.